NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Time Series Forecasting:

An Application to Balance Sheet

Federica Vieira Y de Araoz Murillo

Internship report presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

TIME SERIES FORECASTING: AN APPLICATION TO BALANCE SHEET

by

Federica Vieira Y de Araoz Murillo

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

Advisor: *Prof Doutor* Flávio Luis Portas Pinheiro

External Advisor*: Dra.* Vera Lúcia Cardoso Santos

November 2021

Abstract

The Internal Capital Adequacy Assessment Process (ICAAP) provides a qualitative and quantitative assessment of capital risks to which banking institutions are exposed to in their activity. Caixa Geral de Depósitos (CGD) is a relevant player in the Portuguese banking system, and as such it has to perform an ongoing review of ICAAP exercise to evaluate its ability to identify, assess, mitigate and report on its risks.
In order to properly quantify all the risks the institution is exposed to, several models need to be developed to help estimate the amount of capital that is needed to cover potential unexpected losses arising from each type of risk. Given the European and Portuguese guidelines these models also have to comply with certain requirements defined by Banco de Portugal, European Central Bank (ECB) and European Banking Authority (EBA) regarding ICAAP exercise.

One of the risks CGD is exposed to is the risk of an unfavourable evolution of the main credit items in its Balance Sheet and as such, it is necessary to estimate the evolution of certain credit items (in terms of their volumes and spread rates). These estimations are needed for relevant segments such as housing credit, consumer and other credit, public sector credit, real estate activities credit, non-financial corporate credit and term and sight deposits. To estimate the evolution of these balance sheet items, a robust and reliable methodology must be applied, so that it can truly help strategic decision-making process over a horizon period of three years and the appropriate amount of capital can be allocated.

At CGD, Balance Sheet credit volumes and spread rates had been being estimated through multiple linear regressions to which macroeconomic indicators are added as explanatory variables. The problem with this methodology, is that these type of dependent and explanatory financial variables are usually in the form of time series, indicating the existence of correlation between any observation and the previous one, meaning that there is dependence on the past historical information. Applying multiple linear regressions to this type of data leads to poor statistical results and to the non-compliance of all the statistical assumptions linear regressions must respect.

Within this context, the need to turn to a more adequate and robust methodology became more evident and time series forecasting appeared to be the so long needed solution that would allow to reach reliable statistical results.

Time series forecasting is commonly used in economics and finance, denoting a robust technique to predict macroeconomic variables representing

a feasible approach to apply to estimate CGD's main credit volumes and spread rates of the balance sheet. In this project, we investigate the estimation of Balance Sheet credit volumes and spreads rates using time series forecasting aiming to assess the models suitability to quantify the risk of unfavourable balance sheet evolution of the main credit segments.

The models proposed for this purpose, are the Autoregressive Integrated Moving Average with exogenous variables (ARIMAX) models. The results obtained proved to have robust statistical results and high performance, which were verified by analysing residuals statistical behaviour and key performance indicators such as the Mean Squared Error (MSE) and the Akaike Information Criterion (AIC) of the final models selected for each target variable.

Keywords: CGD, ICAAP, balance sheet projection risk, forecasting, time series, ARIMAX

Contents

List of Figures

List of Tables

Acronyms

CGD	Caixa Geral de Depósitos
ICAAP	Internal Capital Adequacy Assessment Process
RMD	Risk Management Direction (Direção de Gestão de Risco in Portuguese)
IAD	Internal Audit Direction (Direção de Auditoria Interna in Portuguese)
MVO	Model Validation Office (Gabinete de Validação de Modelos in Portuguese)
MSO	Macroeconomic Studies Office (Gabinete de Estudos Macroeconómicos in Portuguese)
SPCD	Strategy, Planning and Control Department (Departamento de Planeamento e Controlo in Portuguese)
EBA	European Banking Authority
ECB	European Central Bank
ESRB	European Systemic Risk Board
ARIMA	Autoregressive Integrated Moving Average
SARIMA	Seasonal Autoregressive Integrated Moving Average
ARIMAX	Autoregressive Integrated Moving Average with eXogenous factors
SARIMAX	Seasonal Autoregressive Integrated Moving Average with eXogenous factors
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
DW	Durbin-Watson
ARCH	Autoregressive Conditional Heteroscedastic
AIC	Akaike Information Criterion
MSE	Mean Squared Error
MINIC	Minimum Information Criterion Method
ESACF	Extended Sample Autocorrelation Function Method

SCAN Smallest Canonical Correlation Method

1. Introduction

This report intends to describe the work developed throughout the internship that took place in Caixa Geral de Depósitos (CGD), as well as the knowledge acquired and the main challenges that were found and overcome. In particular, it describes one of the main projects developed, including the methodologies involved and how the outcomes strategically impact CGD. This section covers the introduction to the work carried out during the internship as well as the description of the professional environment where the internship took place will be covered.

The Model Development Unit (MDU) from the Risk Management Direction (RMD) at CGD aims to develop statistical methodologies to assess and mitigate the different risks to which the banking activity is exposed. One of the themes addressed and which will be the focus of the work presented in this document, is the need to find methodologies to forecast the main items on the institution's balance sheet, namely the evolution of credit volumes and spreads associated with the main products regarding the core business of the bank.

The report focuses on the balance sheet forecast model which makes it possible to simulate different scenarios using macroeconomic variables as explanatory variables, providing a way to stress these components and assess the impact of different macroeconomic scenarios on financial and risk ratios. Indeed, the methodology adopted to build this econometric model was based in time series theory, in particular in ARIMAX models that will be further described in more detail.

Time series modelling is one of the most robust modelling techniques that uses historical data in order to predict the future. This forecasting method consists in collecting, analysing and carefully studying a sequence of observations made over a period of time, and afterwards, use it to predict future values. A very relevant feature of this method is that it requires that the data is collected in equal time intervals [Cowpertwait and Metcalfe, 2009].

Nowadays, this forecasting model is renowned and useful to understand and predict several types of phenomena. For instance, it can be used to estimate the weather, the unemployment rate, the gasoline price, and the real-estate prices, among others. Actually, it can be applied for various purposes in the most diverse areas and have a great impact in business decision-making in certain situations [Papacharalampous et al., 2018].

There are several types of forecasting methods that can be applied in univariate time series. However, after doing some research, the ones that were considered more adequate and useful to test were the Auto Regressive Integrated Moving Average (ARIMA) models, which predict future values given the past values [De Gooijer and Hyndman, 2006]. To these types of models, it is possible to introduce two components: seasonality and exogenous variables, which can enhance the performance of the ARIMA model. Therefore, within this type of

models, the Auto Regressive Integrated Moving Average (ARIMA), the Seasonal Autoregressive Integrated Moving Average (SARIMA), the Autoregressive Integrated Moving Average with exogenous variables (ARIMAX) and the Seasonal Autoregressive Integrated Moving Average with exogenous variables (SARIMAX) models were taken into account as viable options to adopt. In fact, the inclusion of exogenous variables is fundamental to enhance the modelling process and its predictions under different macroeconomic scenarios, so the models which include explanatory variables were preferred.

These time series forecasting methods are powerful and usually provide reliable predictions, which afterwards prove to be consistent with the real-life values. However, it is important to notice that one of the concerns with this method is that overfitting[1] occurs and in order to mitigate this problem, we used an out-of-time sample. Hence, this was the approach designated for projecting the evolution of the main balance sheet items (credit volumes and spreads) of CGD's domestic activity based on a given baseline and adverse scenario. Notice that this modelling technique was applied to data regarding the Portuguese banking sector, under the assumption that the trend of the Institution's items follows the same assumptions as the Portuguese financial system.

The method applied for the purpose of projecting balance sheet items before I started my internship at CGD was multiple linear regressions. Although the estimations provided by the regressions were considerably acceptable, this methodology had some recommendations issued by the Internal Audit Direction (IAD)[2], concerning the compliance of all the statistical assumptions that linear regressions must fulfil, that needed to be addressed.

Consequently, and considering the suggestion within the issued recommendations, a more robust methodology was researched and the one that was considered to be more effective and applicable for the purpose of Balance Sheet projections, was time series forecasting, as already mentioned. In CGD's particular case, these estimations are of great importance in helping defining the business strategy and decision-making process of the bank. Considering the robustness and direct practical improvements, this new methodology brought, turned out to be a great learning opportunity with major responsibility to be involved in this project.

1.1 Company Overview

The internship took place at CGD, the largest commercial banking institution in Portugal, having more than 144 years of history and being present worldwide. This institution is owned by the Portuguese government, and when it was established, its main purpose was to receive mandatory deposits constituted by law or court enforcement[3].

In the second decade of the twenty-first century, a financial crisis, emerged in the context of the public debt crises in the Eurozone, started to affect Portugal. Dealing with this crisis

[1]Overfitting occurs when the model adjusts perfectly to the training set and may not adjust so efficiently to real data.

[2]IAD is the department responsible for auditing all the CGD internal processes, as well as assuring its optimization and efficiency.

[3]`https://www.cgd.pt/Institucional/Patrimonio-Historico-CGD/Historia/Pages/Historia-CGD.aspx`

required international help for the country to regenerate. As a result, in 2011 the Portuguese government announced that they addressed a request for financial assistance to the European Commission to guarantee financing conditions for Portugal and its financial system. After that, 'Troika'[4] presented the financial assistance program to Portugal of 78 billion euros.

However, as CGD did not have access to Troika's fund for its recapitalization, the government decided, in June 2012, to make a capital injection of 750 million euros and also subscribe to 900 million euros in bonds convertible into capital. Nonetheless, since the bank approved several high risk credit concessions that led to critical losses (more than 1 billion euros until 2015), it had serious difficulties in re-establishing, having accumulated 3.416 billion euros in losses since 2011 until 2018.

Ten years later, in the first semester of 2021, the bank achieved a net profit of 294.2 million euros. With this results, Moody's raised CGD's rating to Baa3, fitting in the quality investment category[5]. This was an extremely important achievement for the institution's evolution and positioning in the market, which occurs as a result of the progressive reinforcement of the solidity, profitability and quality of assets applied with the implementation of the 2017-2020 Strategic Plan.

Also, in the first semester of 2021, the bank domestic activity reported a volume of business of 128.098 million euros. Notice that the CGD Group has several entities being represented in about 23 countries spread worldwide, having a share capital of over 3.8 thousand million euros. Particularly in Portugal, there are 3.7 million customers, corresponding to 36% of the country's population[6].

Regarding future goals, the banking institution intends to continue innovating and improving the provision of services in the main business area which is retail banking, open new contact channels with customers to facilitate access to services, promote the use of new technologies by customers and employees in order to increase the quality of the provided service, among others.

Nevertheless, at the moment, its main ambition is to become a leader in sustainable financing in Portugal, supporting the transition to an economy of low carbon and financing projects with a social impact on people's lives.

Obviously, CGD has a great diversity of areas working together to reach these goals, and each one of it has a great impact in the company's business. In particular, one of these areas is where the internship took place which was the RMD.

The Risk Management Department is a first-level body in CGD's organic structure, with control functions and whose object is to protect CGD Group's capital, namely through the

[4]Troika is a decision group made up of three elements, the European Commission (EC), the European Central Bank (ECB) and the International Monetary Fund (IMF), whose responsibility is restructuring the economy of countries who are in a severe economic situation, by providing financial loans.

[5]https://www.cgd.pt/Investor-Relations/Informacao-Financeira/CGD/Relatorios-Contas/2021/Documents/Relatorio-gestao-contas-CGD-1Semestre-2021.pdf

[6]https://www.cgd.pt/Investor-Relations/Informacao-Financeira/CGD/Relatorios-Contas/2021/Documents/Relatorio-gestao-contas-CGD-1Semestre-2021.pdf

management of capital and solvency risks, credit, market, liquidity, interest rate of the banking portfolio, operational and non-financial risks incurred by the Group, the interrelations existing between them and ensuring the coherent integration of their partial contributions. Within the scope of its attributions, this department is responsible for the management of transversal exercises such as Risk Appetite Statement, Internal Capital Adequacy Assessment Process, Internal Liquidity Adequacy Assessment Process, Recovery Plan and the Stress Test, as well as the dissemination of the risk culture by the various Group entities.

1.2 Team and Activities

The MDU is integrated in the RMD, and aims, together with the risk specialized areas, to develop and monitor the internal models used in risk management across the CGD Group. For this, it uses forecasting tools adapted to different risk typologies, applying statistical and advanced analytics methodologies.

With respect to monitoring the various models, the work developed includes the evaluation of the adequacy of the models to their purpose and the application of improvements to ensure the minimization of the risks to which they intend to respond. The work that the MDU team does every day can be defined as a constant problem-solving exercise and continuous improvement of risk assessment and management tools, being necessary to consistently study new methodologies and innovate in a world of constant changing with new risks arising and new ways of identifying them.

In short, its mission is to introduce innovation to the process of risk quantifying and produce precise estimations in order to obtain high performance risk management models.

The MDU team is composed by 10 elements, and as a member of the team, my main responsibilities consisted in helping and working together with the team in several projects regarding different types of risks.

Firstly, I started working in the Balance Sheet projections project, which aimed to estimate, under two different macroeconomic scenarios, the evolution of the main balance sheet items (credit volumes and spreads) of CGD domestic activity.

Secondly, I was involved in the development of the Operational Risk Loss Forecast Model, which intended to predict the losses arising from Operational Risk events that can occur within the CGD Group. These events can be, for instance, internal or external fraud, conduct risk, business disruption and system failures, among others. For this project in particular, it was necessary to acquire software knowledge and skills, predominantly in Excel and VBA, in order to carry out a more detailed data analysis that would automatically be displayed according to the intended group entity.

Afterwards, my contribution was required to develop the automatization of an already existent model that was implemented in Excel until that moment, therefore entailing a high risk of potential operational errors. The model automatization was related to real estate and consisted in quantifying the losses for the entity, arising from real estate risk. The purpose

of this project was to replicate the excel process using SAS software[7], which would allow to execute the whole process in a shorter period of time, and also, reduce the chance of incurring in operational risk errors.

Additionally, I helped in the development of an estimation model for Loss Given Default (LGD) for different product types of Credit Leasing Factoring (CLF)[8]. Generally speaking, this model aims, in the event of a credit default, to estimate what would be the loss for this entity, and therefore, give a support in establishing the spread values for the concession of the credits. Within CLF products, projections were calculated for furniture leasing, real estate leasing and consumer credit.

Later on, I worked in the automatization of the Operational Risk Loss Forecast Model process, so that the model owner[9] could independently estimate the losses related to Operational Risk events.

Overall, among all the projects in which I was involved, the most challenging one for which I was deeply and mainly responsible for and which will be described in detail in this report, consisted in reviewing and improving the implemented methodology used for estimating items of CGD's Balance Sheet. This estimation allows for the calculation of the final capital requirements in adverse scenario that is ultimately one of the final purposes of the Internal Capital Adequacy Assessment Process (ICAAP) exercise[10].

The already existent methodology had some recommendations from the IAD, which were necessary to address in order to improve the statistical robustness of the model. Hence, it was decided to use time series forecasting since it is considered to be one of the most effective methods to make this type of predictions. Afterwards, it was essential to learn all the theoretical concepts related to this method to then, put them into practice using the available technologies provided by the company.

Eventually, after executing all the fundamental procedures, the model was developed and the forecast results of the main balance sheet items of CGD were obtained. These outcomes revealed to be of crucial importance in supporting future strategic decision-making of the bank.

1.3 Internship Goals

The main purpose of the internship was to integrate the MDU, from the RMD, and contribute to develop and improve risk models and its inherent processes. The program required participating in various activities, ideally working with all of the team members and covering several risk areas. Most of the tasks involved data analysis, data cleaning, applying

[7]SAS is a statistical software developed in 1960 by SAS Institute. Essentially, it is used for business intelligence, data management, advanced analytics, as well as predictive, descriptive and prescriptive analysis. The SAS software that is mainly used in the MDU team is SAS Enterprise Guide.

[8]CLF was an independent group entity that was recently integrated in CGD.

[9]Model owners are represented by the business areas that identify the need of a model development and that will apply it in their working activities.

[10]ICAAP is an exercise for the banking sector that consists in evaluating the necessary internal capital to cover for unexpected losses.

diverse modelling methods, automating already implemented processes, elaboration of model documentation, among others.

These tasks required having solid skills in programming, as well as expertise in working with a substantially amount of data using advanced analytical approaches. Taking into consideration that my background is strongly related with working with diverse programming languages, and given that the year before integrating this internship I worked with diverse data sets and obtained the main insights from it, my skills turned out to be crucial to have a successful performance. Although it was the first time that I worked with some of the necessary technological tools, having the competence and the rationale of programming, was extremely advantageous to quickly get comfortable with these new tools.

Even though this internship demanded the participation in several projects, the one that required more time and effort to develop was designated as "Balance Sheet projections". Certainly, this responsibility was directly related with the impact of the conclusions obtained.

As it was mentioned previously, this project consisted in reviewing and improving an already implemented methodology used for estimating items of CGD's Balance Sheet. Since the previous approach used for this model was statistical regressions which did not comply with all the necessary statistical assumptions, it was suggested to use another method that proved to be more efficient. Indeed, time series forecasting was the methodology indicated for this purpose. To develop this particular project, it was necessary to perform certain tasks. Primarily, it was essential to study in detail the main concepts of time series forecasting. Also, there are several statistical models that can be used to forecast time series and the theoretical detail of each one had to be deeply understood. The models under consideration were ARIMA, SARIMA, ARIMAX and SARIMAX, and will be further described on section 2.3.1.1. Subsequently, testing these models, required some effort in getting familiar not only with the methodology, but also with the software tools[11] provided for this purpose.

After understanding the purpose of each of the models, and testing the ones that seemed to be appropriate for the target variables, the forecasted values were carefully analysed. This analysis was essential to select the model to be applied and predict the future values of the main balance sheet items. These results are then transmitted to the administrators of the RMD, in order to assist in forthcoming business decisions with a great impact in the strategic future of CGD.

[11]SAS Enterprise Guide was the main software where time series forecasting models were developed, for which it was necessary to get familiar with SAS language. In particular, SAS Data Step was learned during the internship, while SQL was already known but more experience was acquired. Excel was another tool that was used and which was already known, but it required a deepening of the existent knowledge.

2. Theoretical Framework

This section intends to cover the background and technical knowledge essential for the estimation of the balance sheet items of CGD. More specifically, it will describe the ICAAP and Stress Test exercise to which every banking institution is submitted to every year and how they are related to the project developed along the internship. Additionally, the modelling approach used, as well as some other options that could have been used in the estimation will be discussed.

2.1 ICAAP and Stress Test

Every financial crisis has a strong impact in society, and if the banking sector cannot prevent the shock, it is important to at least, be aware and prepared for its occurrence. When the banking sector capital is inadequate and has a low quality, the depth and the severity of the financial shocks are normally amplified. In the recent financial crisis, credit institutions were forced to re-establish their capital bases when the situation was critical to do so. Additionally, many risks were not properly covered by a quantifiable amount of capital due to deficiencies regarding risk identification and assessment by credit institutions.

Therefore, it is of extreme importance to increase the resilience of the credit institutions in stressful moments, through improvements in their Internal Capital Adequacy Assessment Process (ICAAP), which include exhaustive stress tests and capital planning. The formalization of the policy to be followed for the maintenance of capital levels appropriate to the business risk strategy is done with the application of the ICAAP exercise.

ICAAP is a crucial exercise for banks, which has to follow specific scenario requirements defined by Banco de Portugal. Also, all the results and findings resulting from the exercise of every Portuguese financial institution have to be reported to Banco de Portugal[1].

There are two macroeconomic scenarios that ICAAP considers, designated as baseline scenario and adverse scenario. The baseline scenario consists in a stable economic scenario based on historical data and predictions made by the national central banks. The adverse scenario describes the change of both financial and economic variables assuming a worst-case scenario incorporating the most probable shocks associated with the current banking system in the European Union.

On the whole, ICAAP consists in having a qualitative and quantitative assessment of the risks to which the institution is exposed in its activity, measuring the internal controls and their effectiveness in mitigating exposure to these risks and the simulation of a set of adverse

[1] `https://www.bportugal.pt/comunicado/banco-de-portugal-publica-instrucao-relativa-ao-processo-de-autoavaliacao-da-adequacao-do`

scenarios with influence on solvency[2]. All the findings are reported to the upper levels of the hierarchical chain to help enhancing future decisions that may impact the company's business.

Alternatively, with the aim of developing a unique secure and competent market for financial products in the European Union, the European Banking Authority (EBA) conducted the Stress Test exercise. In order to reach this goal, EBA regulates and supervises the banking sector in all EU countries by establishing a set of rules for the banking institutions to comply with. Particularly, the European Central Bank (ECB) together with the European Systemic Risk Board (ESRB) and the European Commission, are responsible for determining all the components, methodologies, scenarios and statistical tests that the banking institutions must comply with, in this exercise. The outcomes obtained assist in the identification of potential risks or vulnerabilities that each institution faces.

There are several deliverables regarding ICAAP exercise to which CGD must comply with, and the MDU in particular, develops several models which are validated internally to furtherer fulfil the ICAAP requirements. More specifically, the main project developed along the internship concerns the estimation models necessary to provide three year projections of the main balance sheet credit items under two economic scenarios.

In particular, time series methodologies are widely used for the study of macroeconomics with applications in the financial area. The models that were developed for the purpose of the ICAAP exercise allowed to estimate quarterly growth rates evolution for certain balance sheet items for a time period of three years, thus supporting the calculation of risk in different scenarios required for filling out the ICAAP report. Since Stress Test defines a set of good practices for the entities to follow in the exercise, these guidelines were also taken into account for the development of the used methodology.

2.2 Literature Background

The approach used before I started my internship at CGD to estimate growth rates for the main credit items of CGD's balance sheet, consisted in applying multiple linear regressions with macroeconomic indicators as explanatory variables. These credit items are related to several business segments such as housing credit, consumer credit, term and sight deposits, real estate activities, public sector, and non-financial corporations. In the context of evaluating the continuous improvement of the procedures and methodologies used to estimate the ICAAP indicators, all methodologies used to quantify the several types of risks are subject to independent validation and auditing by third defence control areas within CGD. As such, some recommendations were issued by IAD regarding the multiple linear regressions approach that needed to be addressed.

Specifically, the regressions used for the projection of credit volumes and spread rates, although statistically relevant, represented, in some cases, relationships with poor economic

[2] https://www.eba.europa.eu/sites/default/documents/files/documents/10180/1645611/6fa080b6-059d-4b41-95c7-9c5edb8cba81/Final%20report%20on%20Guidelines%20on%20ICAAP%20ILAAP%20%28EBA-GL-2016-10%29.pdf?retry=1

rationale, and no documentation had been produced to substantiate the economic rationale established for the use of the estimated regressions. Also, the regressions were chosen every year depending on their statistical performance and adherence to business expectations, meaning that the models used each year were always different and there was no stabilization of the model applied to predict each balance sheet item.

Additionally, there was also a concern regarding the fact that linear regressions were not sufficiently statistically robust due to two main reasons. The first was that there was no evidence of a prior analysis of variables that could describe credit and deposits volumes and spreads evolution trends, and the process involved evaluating a series of independent variables without identifying first those that were potentially explanatory for the target variables. Secondly, the statistical models were not suitable for the historical data, leading to under or overestimation and opposite signs in the regression coefficients, causing unreliable estimates. Besides, the regressions did not fulfil all the statistical assumptions that are required with special attention to the lack of residuals independence meaning that the models presented self-correlation in the estimation errors.

Since most macroeconomic data usually comes in the form of time series [Cochrane, 2005], meaning that there is dependence in data over time and one historical observation is correlated with the previous observed value, this type of forecast represents a viable approach to be considered for the projection of CGD's main balance sheet items. Indeed, time series has been widely used to forecast future values in the most diverse fields [Chatfield, 2000]. In particular, forecasting time series has been targeted as an efficient method to predict macroeconomic variables since it is considered to be an important topic in business, economics and finance [Sparks and Yurova, 2006; Taylor, 2008].

A time series can be defined as a set of data points collected with equal time intervals over a determined period of time. These observations are then used to forecast future values, that is, the historical data is used to predict future data points, under the main assumption that the past influences and helps to predict the future [Brockwell and Davis, 2009].

Effectively, there are several forecasting models that can be used for this purpose. Nonetheless, since the time series on issue is of the univariate type, the types of models that are mainly applicable for the purpose are the Integrated Auto-Regressive Moving Average (ARIMA), the Long Short-Term Memory (LSTM), and the Generalised Autoregressive Conditionally Heteroscedastic (GARCH) models.

LSTM are a specific type of Artificial Neural Networks (ANN), which are non-linear predictive models that learn through training and resemble biological neural networks in structure [Hochreiter and Schmidhuber, 1997]. LSTM is a Recurrent Neural Network (RNN) based architecture, being capable of storing data for longer periods of time [Sak et al., 2014]. Although it is usual for neural networks to provide better performance results [Kohzadi et al., 1996; Mombeini and Yazdani-Chamzini, 2015; Ho et al., 2002; Prybutok et al., 2000; Catalão et al., 2007], these models are computationally more expensive, require wider ranges of data, have more input parameters, and are less interpretable than simpler traditional models [Namazian et al., 2018; Chakraborty et al., 2019].

Regarding time series forecasting performance, some studies were made to compare the re-

sults of LSTM and ARIMA models. Although, most of them reached better results applying LSTM, the difference was not significant. It was concluded that ARIMA work better in the short-term [Alkali et al., 2019], while LSTM is more appropriate for long-term [Muzaffar and Afshari, 2019; Agrawal et al., 2018]. Also, the outcomes obtained were that it is unworthy to apply such a complex model as LSTM is, while ARIMA is a precision model that can provide identical and reliable results with less complexity involved [Azari et al., 2019; Barman, 2020; Cracan, 2020; Liu et al., 2021].

In the economics and financial fields, GARCH has also been seen as a potential model to forecast time series [McAleer and Da Veiga, 2008; Frimpong and Oteng-Abayie, 2006]. This is a type of ARMA which aims to deal with heteroscedasticity in residuals, that is, it was designed to deal with irregular error variances [Lamoureux and Lastrapes, 1990; Wang et al., 2005]. Thus, when an ARIMA model does not present constant standard deviation in errors, this heteroscedastic model provides the ability of solving this issue. In fact, GARCH models are applied in order to capture past volatility and then predict future volatilities [Sparks and Yurova, 2006]. However, tuning the parameters and optimizing the model, may be a considerable challenge to deal with [Wu and Philip, 2005].

In the particular case of the balance sheet target variables and since time series data did not contain such irregularities, meaning that the variance of the errors was constant, the application of GARCH was not required. Regarding LSTM, it was considered that the results would not provide such transparency as the ones obtained with ARIMA, being an important aspect, given the data aimed to predict and the need to easily interpret the model results, as well as explain them to the CGD board. Since the guidelines of EBA Stress Test were aimed to be accomplished, it was important to be able to explain the model, as well as fulfil its main components and statistical tests. These requirements are possible to achieve with ARIMA, given that neural networks are exhaustive methods which do not allow the adjustment of the model, and are more vulnerable than ARIMA in forecasting.

Thus, after taking into consideration all the possible approaches, it was decided to test the ARIMA method and evaluate its performance. This model is extremely dominant when predicting short-term data and, in particular, in the economics area [Almasarweh and Alwadi, 2018]. Also, following the analysis of several studies comparing the main forecasting models, the outcomes were very clear, that is, ARIMA is one of the most powerful methods to forecast time series, providing high accuracy results [GERUNOV, 2016; Dinardi, 2020; Leszko, 2020]. Although there were other models that proved to have a very good performance, sometimes even better than ARIMA, they were not as appropriate for the data series collected, nor the data aimed to predict.

Hence, this report will be focused essentially in the application of the different variants of ARIMA to estimate the main items of CGD's balance sheet for ICAAP exercise.

2.3 Methodology

This section covers the theory required for the implementation of time series forecasting in the estimation of CGD's Balance Sheet credit volumes and spread rates, as well as some

output examples (applied in SAS) that needed to be analysed for this modelling process.

2.3.1 Model and Analysis

The models under study for this project were based on time series forecasting theory. There are two fundamental processes to develop a time series model. First, time series analysis that consists in analysing the series using several methods to understand how data changes over time, checking if they have patterns such as trend, seasonality, stationarity, among others. Then, the identification and estimation of the model involves identifying the best model and the estimation parameters to be applied in order to reach the best results.

There are several statistical models that can be used to forecast time series. In the following subsections, the theoretical detail of the ARIMA, ARIMAX, SARIMA and SARIMAX models are presented.

By analysing a time series, it is generally intended to achieve two objectives, namely, modelling and forecasting future values, which are closely related. The first objective is to find a model that takes into account the existing relationships between observations, allowing the description of the time series. The second objective concerns the prediction of future values, which is achieved by applying a robust model that has been tested on past values.

Notice that there are several tests that have to be done in order to choose the best model, which will be further detailed. Also, after applying the model there are various procedures that need to be done to evaluate which is the best model for each specific target, which will also be further detailed.

In particular, the validation of the inclusion of the independent variables, the statistical tests and the performance indicators will be further described on some of the following subsections.

2.3.1.1 Time Series Forecasting Models

In this section, the models that were considered for the implementation of time series forecasting will be described in detail. Namely, the Integrated Autoregressive Moving Average (ARIMA), the Seasonal Autoregressive Integrated Moving Average (SARIMA), the Autoregressive Integrated Moving Average with eXogenous factors (ARIMAX) and the Seasonal Autoregressive Integrated Moving Average with eXogenous factors (SARIMAX) models.

2.3.1.1.1 ARIMA Integrated Autoregressive Moving Average (ARIMA) models, are linear statistical models for analysing time series.

The term autoregressive is associated with autoregressive models ("autoregressive" - AR). The AR model is a linear regression of the current value of the series over one or more of the previous values of the series.

The designation "integrated" is explained by reconstructing the original series from the differenced series, through an integration operation or a recursive sum.

Finally, the denomination of moving averages (MA) is related to the fact that this model is built as a linear regression of the present value of the series on the random perturbations of

one or more previous values of the series.

In theory, ARIMA models (p, q, d) are the most general class of models for predicting a time series that, not being stationary, can be stationary through differentiations.

A non-seasonal ARIMA model is classified as a model [Brockwell and Davis, 2009]:

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)a_t \tag{2.1}$$

Where:

- p measures the autoregressive order of the model;
- d is the number of times that the series was differenced in order to be stationary;
- q measures the moving average order;
- a_t is the white noise process;
- $(1-B)^d$ is the differencing operator or in other words, the d-degree integrated part;
- $\theta_q(B)$ is the moving average process;
- $\phi_p(B)$ is the autoregressive process;
- $(1-B)^d Y_t$ is the backshift notation[3].

The terms p, q and d are all integers greater than or equal to 0.

2.3.1.1.2 ARIMAX The integrated autoregressive model of moving average with exogenous variable (ARIMAX) is another statistical model that intends to analyse time series. This can be seen as an ARIMA, but with the difference that ARIMAX has exogenous inputs.

An ARIMAX model is classified as follows [Victor-Edema and Essi, 2016]:

$$\phi_p(B)(1-B)^d Y_t = \varphi_q(B)X_t + \theta_q(B)a_t \tag{2.2}$$

Where:

- X_t is the exogenous variable at time t.

[3]For instance, $(1-B)Y_t$ corresponds to the first difference, $(1-B)^2 Y_t$ is the second difference, in general, a d^{th} order difference can be written as $(1-B)^d Y_t$.

2.3.1.1.3 SARIMA The integrated seasonal moving average autoregressive model (SARIMA) is composed of an ARIMA model, with the insertion of a component that represents the seasonality present in the time series.

The SARIMA model is described using the following equation [Cools et al., 2009]:

$$\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_q(B^s)a_t \tag{2.3}$$

Where:

- s is the seasonal component of the model;
- $\Phi_p(B^s)$ is the seasonal autoregressive order;
- $\Theta_q(B^s)$ is the seasonal moving average order.

2.3.1.1.4 SARIMAX The seasonal moving average integrated autoregressive model with exogenous variable (SARIMAX) is based on the SARIMA model, where the only distinction is that SARIMAX has exogenous inputs.

The following equation represents the SARIMAX model [Arunraj et al., 2016]:

$$\phi_p(B)\Phi_p(B^s)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_q(B^s)a_t \tag{2.4}$$

It is important to notice that for the balance projection to be carried out according to different scenarios, the estimation methods must include independent variables, and the options that allow it are ARIMAX and SARIMAX type models.

2.3.1.2 Implementation Requirements

In order to realise which of these four approaches would be the best with the purpose of predicting credit items, it was essential to analyse the collected series. This implied evaluating if trend, seasonality or stationarity were present in the target series and therefore, understand the most relevant statistics associated to the data. The time series analysis done throughout the project is described along the following subsections.

2.3.1.2.1 Stationarity A time series will be stationary if its statistical properties do not change over time. The stationary series has constant mean and variance over time, and the covariance between lagged values of the series depends only on this distance, that is, on the temporal distance between the two periods.

In Figure 2.1, graphical observations for stationary and non-stationary series are represented:

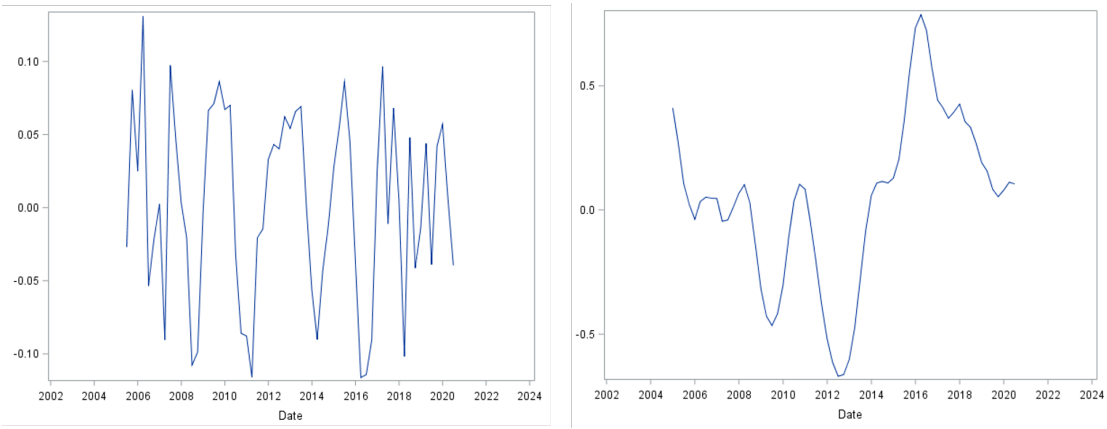

Figure 2.1: Stationary vs Non-stationary Time Series

The stationarity of the series of dependent variables was evaluated through graphical observation, as well as through the evaluation of the respective autocorrelation functions and the Augmented Dickey-Fuller (ADF) test, detailed below.

Autocorrelation Function The autocorrelation function (ACF) is a measure of the correlation between observations of a time series that are separated by k time units (k lags) and can be used to verify the stationarity of a given series. A series is said to be non-stationary if the autocorrelation function decreases slowly, between lags.

The graphs shown in Figure 2.2 illustrate the autocorrelation function for a stationary (left) and a non-stationary (right) series.

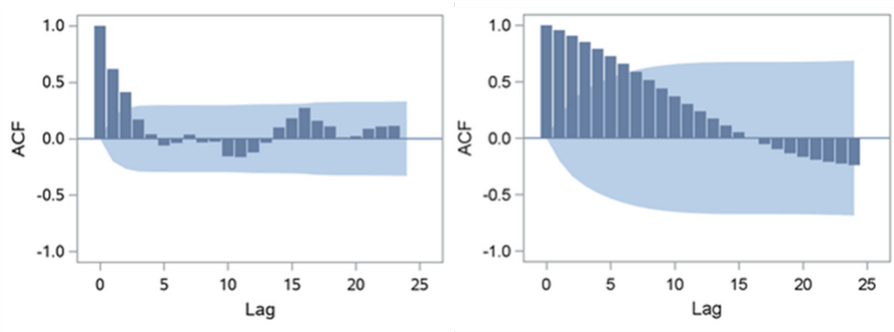

Figure 2.2: Autocorrelation Function for a Stationary vs a Non-stationary Series

Augmented Dickey-Fuller The Augmented Dickey-Fuller (ADF) test is one of several existing tests for unit roots in time series. As an augmented version of the original Dickey-Fuller test, it was designed to be applied to more complex time series models. Specifically, this test is carried out in order to verify the stationarity of the series, that is, it analyses whether or not a series contains a unit root.

Thus, the following regression is studied [Institute, 2012]:

$$y_t = \alpha y_{t-1} + \epsilon_t \tag{2.5}$$

Where α corresponds to the autoregressive coefficient of the time series.

This test considers the following statistical hypotheses:

- H0: $\alpha = 1$, when the series is non-stationary;
- H1: $|\alpha| < 1$, when the series is stationary.

The test statistic is as follows:

$t_{obs} = \frac{\hat{y}}{\hat{\phi}_y}$

Where:

- $\hat{y}$ is an estimate for y;
- $\hat{\phi}_y$ is an estimator for the standard deviation of the error y.

Decision Rule:

- If $t_{obs} < t_{crit}$, the null hypothesis (H0) is rejected, which means the series is non-stationary.
- If $t_{obs} > t_{crit}$, the null hypothesis (H0) is not rejected, so the series is stationary.

For a t_{crit} of 5% (0.05), the null hypothesis (H0) is not rejected, but in case the *p-value* < 0.05, H0 is rejected and the series is stationary. The Tau test statistic is most commonly used to decide whether or not to reject the null hypothesis of non-stationarity (or existence of a unit root) [Institute, 2012]. Thus, the interpretation was performed observing the *p-value* of Tau for the Zero Mean, which is the indicated model.

Figure 2.1 shows examples of ADF test results. In the first case the series is non-stationary since the *p-value* is $0.4118 > 0.05$, and in the second the conclusion is that the series is stationary since the *p-value* < 0.0001.

Table 2.1: ADF Test for a Non-stationary vs Stationary series

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-7.4088	0.0562	-0.70	0.4118		
	1	-25.0006	0.0001	-1.37	0.1578		
Single Mean	0	-7.9372	0.2050	-0.74	0.8275	0.73	0.8850
	1	-25.2189	0.0014	-1.39	0.5826	1.44	0.7071
Trend	0	-8.6346	0.5077	-0.82	0.9584	1.53	0.8711
	1	-23.8250	0.0183	-1.36	0.8622	1.95	0.7893

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-93.9009	<.0001	-13.39	<.0001		
	1	-226.337	0.0001	-8.93	<.0001		
Single Mean	0	-94.1103	0.0006	-13.34	0.0001	88.96	0.0010
	1	-228.188	0.0001	-8.93	0.0001	39.95	0.0010
Trend	0	-94.6712	0.0001	-13.47	<.0001	90.79	0.0010
	1	-234.625	0.0001	-9.10	<.0001	41.72	0.0010

2.3.1.2.2 Trend The trend consists of a certain long-term behaviour of a time series, that is, when a series shows constant increases or declines in successive periods of time, it is possible to conclude that this type of variation is present. The trend is considered a component of time series and prevents its stationarity. Thus, when evaluating stationarity, it is also possible to draw conclusions regarding the trend of the series.

In order to remove the trend, that is, ensure that the series is stationary, the differentiation process must be applied. This method consists of making successive differences in the series until it becomes stationary.

The first difference can be defined as [Hyndman and Athanasopoulos, 2018]:

$$\Delta Y_t = Y_t - Y_{t-1} \tag{2.6}$$

Where t represents the current observation of the time series. To apply the second differencing the following formula is applied:

$$\begin{aligned} \Delta^2 Y_t &= \Delta[\Delta Y_t] = \Delta[Y_t - Y_{t-1}] \\ \Delta^2 Y_t &= Y_t - 2Y_{t-1} + Y_{t-2} \end{aligned} \tag{2.7}$$

Generally speaking, the n$^{\text{th}}$ difference of Y_t is calculated this way:

$$\Delta^n Y_t = \Delta[\Delta^{n-1} Y_t] \tag{2.8}$$

Thus, if after applying the first difference the trend is not removed, the second difference will be tested. If the trend persists, the differentiation must be successively applied to ensure that the series is no longer trending and is stationary.

Figure 2.3 illustrates how trend can be observed in a graphical visualization.

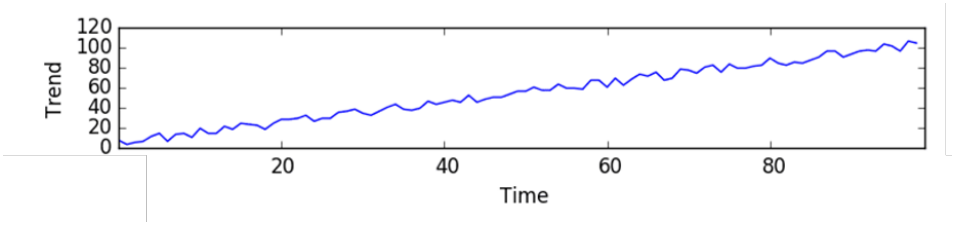

Figure 2.3: Time Series with presence of Trend[4]

2.3.1.2.3 Seasonality A time series is said to be seasonal when the phenomena that occur during time are repeated in each identical period of time.

In Figure 2.4 is possible to observe the occurrence of seasonality. A practical example of the presence of this phenomenon can be a company that has sunscreens for sale, being expected to sell more in summer than in other times of the year.

[4] `https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality`

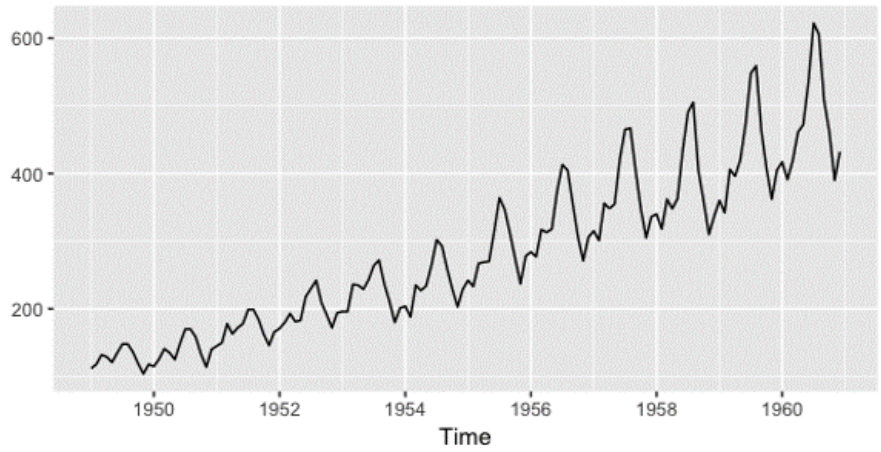

Figure 2.4: Time Series with presence of Seasonality[5]

2.3.1.2.4 Identification and Estimation of the Model In order to identify the order of an ARIMA model, that is, how to establish the base equation, identifying p, d and q, it is necessary to use the autocorrelation functions (ACF) and the partial autocorrelation functions (PACF). In particular, it is through this process that the model to be used is selected.

The d represents the number of times the series has been differentiated in order to become stationary. The p measures the autoregressive order of the model. If the time series has an autoregressive order of 1, called AR(1), then only the first coefficient of the partial autocorrelation should be significant. However, if you have an AR(2), both the first and second coefficients of the partial autocorrelation should be considered significant. Note that they can be positive or negative, what is verified is whether they are statistically significant. Generally, the partial autocorrelation function will have significant correlations up to lag[6] p, and will decrease to values close to zero after lag p.

[5] `https://medium.com/@sangarshananveera/time-series-forecasting-part-1-d523bcc3acbf`

[6] The lag corresponds to the number of periods in the past counting from the actual observation. For instance, if the correlation at lag 0 is being measured it means that the correlation with the actual observation is being calculated (which will be 1 since the observation will be perfectly correlated with itself). While if the correlation at lag 12 is being measured, it means that the correlation between the actual observation and 12 observations before is being calculated.

Table 2.2: Partial Autocorrelation Function

Partial Autocorrelations

Lag	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5
1	0.37064	. |*******
2	0.15912	. |***
3	0.10330	. |**
4	0.04939	. |*.
5	-0.07279	.*| .
6	0.00433	. | .
7	0.01435	. | .
8	0.06815	. |*.
9	0.08346	. |**
10	-0.02903	.*| .
11	-0.03996	.*| .
12	-0.07539	**| .
13	-0.08379	**| .
14	-0.03419	.*| .
15	-0.02101	. | .
16	0.01950	. | .
17	-0.00768	. | .
18	0.01681	. | .

Table 2.2 shows an example of a partial autocorrelation function. It can be seen that the first two (or perhaps three) autocorrelations present in the table are statistically significant (according to the correlation values greater than 0.05 and the number of *[7]). This indicates that it will initially be an AR(2) or AR(3) model.

The q measures the moving average order (MA). In order to take into account which orders should be considered, the autocorrelation function is analysed. If the time series is a moving average of order 1, therefore MA(1), only a significant correlation coefficient should be seen in lag 1. This is due to the fact that the process MA(1) has only 1 period memory temporal. However, if there is a time series MA(2), two autocorrelation coefficients must be significant, in lag 1 and 2, since MA(2) has these two periods in memory. Normally, for a series that is MA(q), the autocorrelation function will have significant correlations up to lag q, with a decline to values close to zero after lag q.

In the example shown in Table 2.3, since lag 4 appears to have an abrupt drop to zero, the MA(4) model can be considered. Thus, initially an ARIMA(2,1,4) is estimated, where the 1 indicates that the series will have been differentiated once to become stationary.

[7]The * represent the calculated correlation, for instance, on lag 1, the correlation is of 0.37, and the * are representing that value.

Table 2.3: Autocorrelation Function

Lag	Covariance	Correlation	Autocorrelations -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
0	0.341391	1.00000	|**********|
1	0.126532	0.37064	|******* . |
2	0.093756	0.27463	|***** . |
3	0.079004	0.23142	|***** . |
4	0.062319	0.18254	|**** . |
5	0.021558	0.06315	|*. . |
6	0.020578	0.06028	|*. . |
7	0.018008	0.05275	|*. . |
8	0.029300	0.08583	|** . |
9	0.040026	0.11724	|** . |
10	0.020880	0.06116	|*. . |
11	0.010021	0.02935	|*. . |
12	-0.0071559	-.02096	| . . |
13	-0.026090	-.07642	**| . . |
14	-0.031699	-.09285	**| . . |
15	-0.032960	-.09654	**| . . |
16	-0.023544	-.06897	. *| . . |
17	-0.021426	-.06276	. *| . . |
18	-0.0084132	-.02464	. | . . |

Another way to determine the p (AR) and q (MA) orders of the models is through the MINIC (Minimum Information Criterion Method), ESACF (Extended Sample Autocorrelation Function Method) and SCAN (Smallest Canonical Correlation Method) methods, whose options can be directly introduced in the ARIMA procedure, used in the SAS software for model estimation. Table 3.7 represented in 3.2.4 illustrates an example of the SAS output relating to the introduction of these options in the model estimation query.

With the SAS results, it is possible to see the combinations of parameters that are associated with a lower BIC (Bayesian Information Criterion) value, since the lower the BIC value, the better the fit of the model to the data. Thus, to determine the order of the models, some of the combinations of parameters provided by the 3 methods were considered.

2.3.1.2.5 Independent Variables Selection Methods Since the intention is to obtain models that allow the inclusion of independent variables to which certain scenarios can be applied, it is necessary to select the variables that allow to help predict the target variable. For this purpose, the Stepwise method was used, which allows a first selection of the explanatory variables that potentially help to predict the dependent variable, and the results of the models based on this method of variable choice were analysed. Additionally, and since the best models are not always those whose variables were selected using the Stepwise method, explanatory variables were also manually selected based on their correlations with each other and with the respective target. Both methods are described below.

Stepwise Method The Stepwise method consists of an iterative construction of a regressive model that involves the selection of independent variables to be used in the final model. This implies including or excluding variables from the model and testing whether they are statistically significant at each iteration performed.

Effectively, it is a process where the variable that enters one step can leave in the next ones and the variable that leaves one step can re-enter the next ones. In fact, this method is often used in models that include exogenous variables such as ARIMAX or SARIMAX. Thus, in some versions of the application of these models, this method was used to make a first selection of the explanatory variables that would help to predict the target variable.

This process adds and removes variables at each iteration, and ends as soon as there are no more variables to introduce or eliminate from the model. Table 2.4 illustrates an example of the last variable selection step produced by the algorithm.

Table 2.4: Stepwise Method SAS Output Example

Note: Model building terminates because the effect to be entered is the effect that was removed in the last step.

Summary of Stepwise Selection								
	Effect							
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Effect Label
1	_1dif_Euribor3M_Var_		1	1	7.1888		0.0073	_1dif_Euribor3M_Var_1Y
2	_1dif_PT_GER_10Y_Var		1	2	7.3136		0.0068	_1dif_PT_GER_10Y_Var_1Y
3	_1dif_TX_JR_habit_EU		1	3	7.4993		0.0062	_1dif_TX_JR_habit_EURO
4	_1dif_indice_prd_con		1	4	3.1677		0.0751	_1dif_indice_prd_const_Var_Lag1Y
5		_1dif_indice_prd_con	1	3		3.0685	0.0798	_1dif_indice_prd_const_Var_Lag1Y

The analysis of the results obtained in the Stepwise method consists firstly in verifying whether the explanatory variables returned from this procedure are statistically significant ($p-value < 0.05$), and if not, the reestimation is carried out with the variables that are, until all of them are statistically significant. In the example above, only the first three variables returned by the algorithm would be considered, since the last one is not significant and is eventually removed.

Manual Selection Methods (Correlation Analysis) Apart from the Stepwise method, the results of models with selected independent variables were also analysed based on their individual correlations with the target to be predicted and based on the correlations between them. For a given model, it is intended that the respective independent variables have a significant correlation with the target variable (ideally above 50%) and that they are poorly correlated with each other (correlations below 50%) in order to prevent the existence of multicollinearity in the models.

The evaluation of correlations was based on Pearson's correlation and the existence or not of multicollinearity in the model was evaluated using the VIF test.

- **Pearson Correlation** Pearson's correlation consists in detecting the existence of correlation between variables. That is, it allows measuring the relationships between variables. This correlation can be either positive or negative. In other words, in the case of being positive, it means that when one of the variables increases, the other also increases. If there is a negative correlation, when one of the variables increases, the other decreases, being therefore inversely proportional.

As it can be seen in Figure 2.5, the graph on the left represents a negative correlation between two variables. The middle view has no correlation. Finally, the graph showing a positive correlation between two variables is presented.

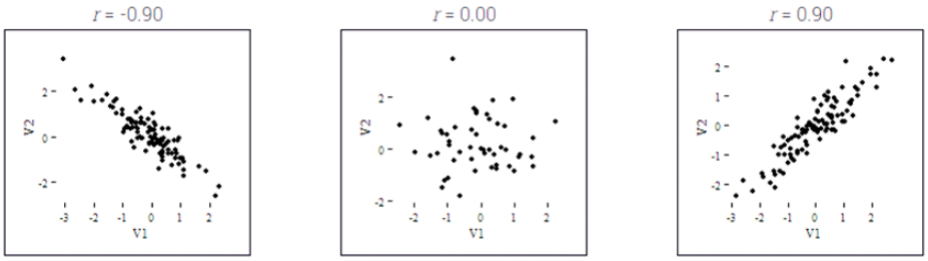

Figure 2.5: Negative, Absence and Positive Pearson Correlation correlation[8]

Hence, the Pearson coefficient can vary between -1 and 1. If the value is equal to -1, it means that there is a perfect negative correlation, being visually very similar to the image on the left, which has a value of -0.9. If there is no correlation, the visualization will be similar to the image in the middle, which has a coefficient equal to 0. Conversely, when there is a perfect correlation, it means that the coefficient will be equal to 1, being graphically identical to the image on the right. Note that Pearson's correlation is only capable of detecting linear relationships, and therefore, if there are non-linear correlations, Pearson's coefficient will not be able to detect them.

In particular, when the Pearson coefficient between the dependent and independent variables is greater than 0.5 or less than -0.5, the explanatory variables are selected to be tested in the model. However, there may also be cases in which variables with correlations below 50% are selected if it is found that the correlations of the independent variables with the target are not generally high and/or if it is understood that a certain independent variable has a coherent theoretical trend to predict the dependent variable, thus ensuring that the models have a robust economic rationale.

- **Variance Inflation Factor (VIF) Multicollinearity Test**

  Multicollinearity occurs when, in a regression, two or more independent variables are correlated with each other. In particular, it is a problem in the adjustment of the model that can impact the estimation of the parameters since it is not possible to distinguish between the individual effects of the independent variables on the dependent variable (target).

  One of the tests that can detect the existence of multicollinearity is the Variance Inflation Factor (VIF).

  The main diagonal elements of $(X'X)^{-}1$ are also useful for detecting multicollinearity. The j^{th} element of the main diagonal $(X'X)^{-}1, C_{jj}$ can be written as:

  $C_jj = (1 - R_j^2)^{-}1, \quad j = 1, ..., p.$

  Where:

[8] `https://libguides.library.kent.edu/spss/pearsoncorr`

- R_j^2 is the coefficient of determination of the regression of X_j on the other explanatory variables;
- C_jj is the variance inflation factor and another notation used is VIF_j.

Therefore, VIF_j is given by:

$VIF_j = \frac{1}{1-R_j^2}$

Where VIF_j measures how much the coefficient's variance $\hat{\beta}_j$ is inflated by its collinearity. If R_j^2 is close to 1, it means that there is a high correlation between the variable X_j and the other, so $1 - R_j^2$ will be close to 0 (zero), and consequently, VIF assumes a high value. As a result, a value greater than 10 indicates that multicollinearity may be influencing least squares estimates.

2.3.1.3 Residuals Analysis

After applying the previously referenced models, it is essential to verify if they pass the main residuals evaluation tests. The tests considered for this purpose are detailed below.

2.3.1.3.1 Ljung-Box Test The Ljung-Box test consists in evaluating the autocorrelations of the estimated residuals, up to a given lag. This test is used to test the failure of the fit of a time series model, and can be performed for any series that has a constant mean. Additionally, it can be considered as an improvement of the Box-Pierce test.

The following hypotheses are considered:

- **H0**: Model does not show fit failure (residuals are white noise);
- **H1**: Model shows fit failure (residuals are not white noise.

The test statistic is shown below [Enders, 2008]:

$$Q(m) = n(n+2)\sum_{j=1}^{m}\frac{r_j^2}{n-j} \tag{2.9}$$

Where:

- R_j is the estimated correlation of the series at lag j.
- m is the number of lags that is being used.

Decision Rule:

If $Q > X_{1-\alpha,h}^2$, the null hypothesis (H0) is rejected and it is possible to conclude that the model shows significant failure of fit.

$X_{1-\alpha,h}^2$ is the value of the chi-square distribution table with h degrees of freedom and a significance level of α.

If the *p-value* < 0.05, the null hypothesis (H0) is rejected, which indicates that there is no autocorrelation and therefore the residuals are not white noise. Otherwise, if *p-value* ≥ 0.05 the residuals are white noise. Examples are given below in Table 2.5 for the two situations mentioned.

Table 2.5: Ljung-Box Not White Noise vs White Noise Residuals

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	25.46	5	0.0001	-0.521	0.014	0.104	0.168	-0.391	0.268
12	27.30	11	0.0041	-0.060	-0.093	0.020	-0.006	-0.092	0.102
18	33.26	17	0.0104	0.048	-0.167	0.174	-0.081	-0.067	0.119
24	42.77	23	0.0074	-0.075	-0.070	0.188	-0.138	-0.070	0.179

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.42	4	0.2472	0.013	-0.071	0.182	0.013	-0.242	0.059
12	10.15	10	0.4272	-0.066	-0.097	-0.077	-0.171	-0.132	0.098
18	13.23	16	0.6560	-0.003	-0.118	0.079	0.054	-0.080	0.109
24	22.14	22	0.4516	-0.033	-0.096	0.104	-0.121	-0.052	0.233

2.3.1.3.2 Durbin-Watson Test The Durbin-Watson (DW) test allows evaluating whether the errors are independent of each other and is used to detect the presence of dependence (autocorrelation) in the residuals of a regression analysis.

The presence of autocorrelation was tested recurring to the following hypotheses:

- H0: Residuals are independent (do not have autocorrelation);
- H1: Residuals are not independent (they have autocorrelation).

Since e_i is the residual associated with the i$^{\text{th}}$ observation and n the number of observations, we have that the DW test statistic is given by [Brocklebank and Dickey, 2003]:

$$dw = \frac{\sum_{i=2}^{n}(e_{i=1} - e_i)^2}{\sum_{i=1}^{n} e_i^2} \tag{2.10}$$

Thus:

- If $0 \leq dw < dL$ then H0 is rejected;
- If $dL \leq dw \leq dU$ the test is inconclusive;
- If $dU < dw < 4 - dU$ then H0 is not rejected;
- If $4 - dU \leq dw \leq 4 - dL$ the test is inconclusive;
- If $4 - dL < dw \leq 4$ then H0 is rejected.

Where dL and dU are the critical values from the DW table [9].

More generally, the DW test analyses the independence of the residuals. If the $p - value > 0.05$, the null hypothesis (H0) is accepted and the residuals are independent.

[9]The critical values of the Durbin-Watson table were taken from `https://www3.nd.edu/~wevans1/econ30331/Durbin_Watson_tables.pdf`

Table 2.6 presented below, shows the comparison between a DW test result when the residuals are not independent, that is, when there is autocorrelation, and another when the residuals are independent, meaning that there is no autocorrelation.

Table 2.6: DW Test with Dependent Residuals vs Independent Residuals

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	0.4752	<.0001	1.0000
2	1.2935	0.0137	0.9863
3	2.0694	0.6545	0.3455
4	2.5544	0.9818	0.0182

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	1.7355	0.0607	0.9393
2	2.1058	0.5519	0.4481
3	2.0286	0.5002	0.4998
4	2.2835	0.8880	0.1120

2.3.1.3.3 Shapiro-Wilk Test Normality tests are used to ascertain whether the probability distribution associated with a data set can be approximated by the normal distribution. The Shapiro-Wilk test is one of the existing normality tests intended to verify the adherence of a sample to the accumulated distribution function.

This test is used when the number of observations is less than 2000 [Razali et al., 2011]. Consequently, it is the indicated test to verify the distribution of data in this project. This test considers the following hypotheses:

- **H0**: Data follows a normal distribution;
- **H1**: Data does not follow a normal distribution.

The test statistic is defined by:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_{(i)} - x^{-})^2} \tag{2.11}$$

Where:

- $x_{(i)}$ are the ordered sample values ($(x_{(1)})$ is the smallest)
- constant b is determined as follows:

$$b = \begin{cases} \sum_{i=1}^{n/2} a_{n-i+1} \times (x_{(n-i+1)} - x_{(i)}) & \text{if } n \text{ is even} \\ \sum_{i=1}^{(n+1)/2} a_{n-i+1} \times (x_{(n-i+1)} - x_{(i)}) & \text{if } n \text{ is odd} \end{cases} \tag{2.12}$$

Where a_{n-1+1} are constants generated by the means, variances and covariances of the order statistics of a sample of size n from a normal distribution, and a_i are coefficients tabulated in Shapiro-Wilk's Table[10].

[10]These coefficients can be obtained from table D5 from the PDF available in the following website: `https://www.epa.gov/sites/default/files/2015-10/documents/monitoring_appendd_1997.pdf`

Decision Rule:

The null hypothesis (H0) is rejected when $W_{Calculated} < W_{\alpha}$, where W_{α} is the critical value[11] at the significance level α, and therefore it is concluded that the data do not follow a normal distribution.

In Table 2.7, it is possible to observe an example in which the data does not follow a normal distribution (right) and another in which the data follows a normal distribution (left).

Table 2.7: Shapiro-Wilk Test with data that follows vs does not follow a Normal Distribution

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.990075	**Pr < W**	0.9037
Kolmogorov-Smirnov	**D**	0.067081	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.036021	**Pr > W-Sq**	>0.2500
Anderson-Darling	**A-Sq**	0.208859	**Pr > A-Sq**	>0.2500

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.888417	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.145982	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.365765	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2.181896	**Pr > A-Sq**	<0.0050

Since W_{α} is 0.947 ($W0.05$), according to the table, and $W_{Calculated} = 0.888417$ in the table on the right, then $W_{Calculated} < W_{\alpha}$, and the data does not follow a normal distribution.

More generally, if the *p-value* > 0.05, the null hypothesis (H0) is accepted concluding that the residuals follow a normal distribution for a significance level of 0.05. Otherwise (*p-value* < 0.05), the alternative hypothesis (H1) is accepted, concluding that the residuals do not follow a normal distribution.

2.3.1.3.4 Autoregressive Conditional Heteroscedastic (ARCH) Test Engle's (1982) [Engle, 1982] Autoregressive Conditional Heteroscedastic (ARCH) test statistic is the most commonly used test that allows to assess the existence of homoscedasticity/heteroscedasticity [Sjölander, 2011]. Homoscedasticity is verified when the error variance, conditional on the explanatory variables, is constant, which is what is intended to be verified in the residual analysis. Conversely, when the error variance is different for each value, that is, when the errors are dispersed in a different way in a regressive analysis, heteroscedasticity is verified. Especifically, when the error variances is not constant, it may indicate that the ordinary least squares (OLS) estimates are inefficient [Institute, 2014] since the disturbances might be correlated [Safi, 2004].

When the error variance is not constant, the data are said to be heteroscedastic, and ordinary least-squares estimates are inefficient

The following hypotheses are considered:

- **H0**: The error variance is constant, proving the existence of homoscedasticity.

[11] These are the values mentioned in the previous footnote.

[11] OLS estimates the parameters of the model in order to minimize the sum of squared residuals [Green, 2001].

- **H1**: The error variance is not constant, proving the existence of heteroscedasticity.

When the *p-value* < 0.05, the residuals are heteroscedastic, and when this is not the case, the residuals are homoscedastic. Table 2.8 shows an example where the residuals are homoscedastic (left) and another where they are not (right).

Table 2.8: ARCH Test with Homoscedasticity vs Heteroscedasticity

Tests for ARCH Disturbances Based on OLS Residuals				
Order	Q	Pr > Q	LM	Pr > LM
1	0.3530	0.5524	0.3154	0.5744
2	0.5164	0.7724	0.4729	0.7894
3	1.0369	0.7923	0.9627	0.8103
4	3.7193	0.4453	3.2249	0.5209
5	4.0142	0.5474	3.3229	0.6503
6	4.0220	0.6737	3.3234	0.7673
7	4.2583	0.7496	3.3557	0.8503
8	4.8248	0.7761	3.5253	0.8972
9	5.4846	0.7902	3.7389	0.9277
10	5.8094	0.8310	3.9466	0.9497
11	6.3930	0.8459	4.1506	0.9653
12	6.4588	0.8912	4.4063	0.9749

Tests for ARCH Disturbances Based on OLS Residuals				
Order	Q	Pr > Q	LM	Pr > LM
1	15.7385	<.0001	14.6086	0.0001
2	15.8349	0.0004	18.8807	<.0001
3	16.3890	0.0009	23.6346	<.0001
4	21.0578	0.0003	23.6979	<.0001
5	21.9018	0.0005	24.5420	0.0002
6	22.4693	0.0010	24.5547	0.0004
7	22.8288	0.0018	24.5990	0.0009
8	23.1453	0.0032	25.6354	0.0012
9	23.7529	0.0047	26.0555	0.0020
10	24.7635	0.0058	26.4363	0.0032
11	25.2424	0.0084	26.4919	0.0055
12	25.6282	0.0121	26.4930	0.0091

2.3.1.4 Performance Indicators

In order to perform error analysis, there are several metrics that can be used. That is, to optimize the performance of the models, it is necessary to have a validation criterion that quantifies the similarity between the predicted and actual values.

The Akaike Information Criterion (AIC) information criterion and the Mean Squared Error (MSE) were the performance indicators chosen for this purpose.

2.3.1.4.1 Akaike Information Criterion (AIC) The Akaike Information Criterion (AIC) consists of minimizing the following formula [Institute, 2012]:

$$AIC = -2ln(L) + 2k \tag{2.13}$$

Where:

- $ln(L)$ corresponds to the logarithm of the maximum likelihood function of the vectors of the estimated parameters;
- k is the number of independent parameters.

In this way, it is possible to obtain the maximum optimal number of lags to be included in the previously mentioned tests. In case the results differ between criteria, this is the most recommended method and was the chosen one for this process.

2.3.1.4.2 Mean Squared Error (MSE)

The Mean Squared Error (MSE) consists in calculating the mean of the squared errors, that is, the square mean of the difference between the estimated values and the true value of the estimated quantity.

In this way, the MSE is used to determine how the model fits the data or not, and whether the elimination of certain data could benefit the model. Thus, this measure allows choosing the best estimator.

The formula to be applied is the following [Sulasikin et al., 2020]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.14}$$

Where:

- MSE corresponds to the mean squared error;
- n corresponds to the number of data points;
- Y_i corresponds to the number of observed values;
- $\hat{Y}_i$ corresponds to the predicted values.

Figure 2.6 shows a series of pink dots that represent the data, the blue line is intended to fit between the points as best as possible, and the red lines represent the errors, which are the distances between the actual observations and the estimated observations.

The goal is to minimize the mean (MSE), which will provide the line that best fits between all the points. That is, if the result obtained is zero (0), this indicates that the estimator predicts observations with perfect precision and so, the objective is to obtain values as close as possible to 0.

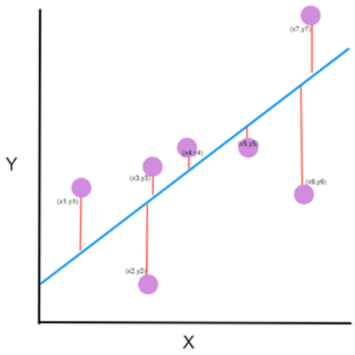

Figure 2.6: MSE Graphical Example[12]

[12] https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93/

3. Projects

Time series models were developed in order to forecast credit volumes and spread rates of CGD. These estimations are indispensable for the annual ICAAP exercise. The aim was to move from multiple linear regressions with annual forecasts to time series with quarterly forecasts, including new explanatory variables and reaching more reliable results.

In this chapter, the workflow and the motivation of the Balance Sheet projection project, as well as the data set worked with and the results obtained in this modelling process will be covered.

3.1 Timeline

Figure 3.1 displayed below presents the global phases of the process of development of CGD's Balance Sheet projection.

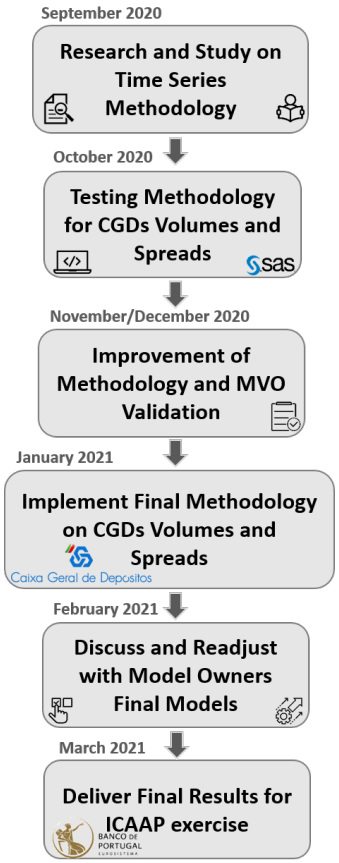

Figure 3.1: Balance Sheet Timeline

The development of this project started in September 2020, where the research and studies

of the theoretical methodology related to time series forecasting began. After having a good knowledge of the application of this methodology, it was tested in an initial stage using SAS software, for CGD's main balance sheet credit items (volumes and spreads), to confirm it would be a robust approach to put in practice for the following ICAAP exercise. Notice that in order to apply the referred methodology, it was essential to collect and compile all the necessary data, including dependent and independent variables, as detailed in section 3.2.2, required for the models, which was a challenging phase of the process.

The whole process to estimate the projections of the dependent variables together with the outcomes obtained, was then delivered to the MVO. After having the positive feedback from MVO for the application of time series forecasting, having just a few recommendations regarding the documentation, the same methodology, having some readjustments if considered necessary, was implemented for CGD's main balance sheet items in order to comply with ICAAP exercise.

Then, once the final results were obtained, the possible models for each of the target variables were discussed with the model owner. This phase required some retesting on the models and several meetings to determine which would be the actual final models to predict each of the credit volumes and spreads included in balance sheet. Nonetheless, in March all the recommendations were addressed and the project was concluded.

Overall, it was a complex and long process which required several retestings, readjustments and reviews on the final models to consider.

3.2 Balance Sheet Projection

The estimation of the main balance sheet items was the key project of my internship. The Balance Sheet Projection models intend to predict the evolution of the main balance sheet credit items (volumes and spreads) of CGD's domestic activity based on a baseline scenario and on an adverse scenario.

These models are useful as they can recreate adverse projections with different degrees of severity that can be used for the different exercises with which the bank must comply with, specifically: serving as a basis for projecting the bank's capital requirements over a horizon period of three years in an adverse scenario based on a review of both the macroeconomic scenarios and the projection of the bank's balance sheet in the baseline scenario.

Throughout the project, there were a few challenges encountered, which were particularly related with the data. Concerning the data quality, there was a difficulty in using internal data due to its high idiosyncrasy, since it is from CGD itself. In other words, the bank has a specific business, which includes particular companies or customers, which does not represent the whole Portuguese population financials. Therefore, it was decided to collect data from Banco de Portugal. When gathering data from Banco de Portugal, the Portuguese financial system is being reflected, and as a result, it is more modelable than if the data was just from CGD. Besides, the market trend is also CGD's trend given its dominant position in the Portuguese market in the main credit products, so it makes sense to use data from the entire universe of the Portuguese banking system.

Additionally, it was significantly challenging to find quarterly historical data for all the required variables. Notice that there were needed at least 30 observations [Davies et al., 2014], since with less than this number, the statistical results are not reliable.

Respecting the process to find the adequate model, there was an extreme difficulty in finding explanatory variables to be included in the model. In order to be included in the model, independent variables must be highly correlated with the target variables and must have the right macroeconomic theoretical tendency of evolution towards the targets, so that the models interpretation is consistent with the expected economic evolution. Even though these conditions can seem limiting, causing various variables to be discarded, they are also essential to have a good forecasting effect and rightful interpretation results.

3.2.1 Motivation

The main motivation for this project's development was to participate in ICAAP and achieve relevant statistical results that could strategically help the institution business. ICAAP (as already described in section 2.1) intends to quantify the capital required to face unpredicted losses. This exercise is extremely important to assist in the predictive ability to anticipate and react to unexpected events. Therefore, it provides a relevant support in CGD's decision-making.

Additionally, the estimations, produced by the methodology that was being used before I started my internship at CGD, were in an annual basis only, and it did not respect the statistical tests requirements. In fact, multiple linear regressions often did not pass the residuals autocorrelation, meaning that sometimes the residuals were correlated and that there were information from the residuals that could be useful in forecasting that was not being used, so this was an issue that needed to be solved.

In order to surpass multiple linear regressions limitations, a more robust methodology such as time series forecasting was found to be a good recommendation. This approach also allowed the prediction for the future three years in a quarterly basis, being more specific than the linear regressions. The implementation of time series models provided important outcomes that can assist in several decisions made for CGD's business future, being extremely relevant for the bank.

3.2.2 Data Sets

This subsection covers the procedures executed to transform the collected raw data into the input data set as well as the detail of the dependent and independent variables used in the Balance Sheet projection.

Regarding the data necessary to apply time series forecasting, these consisted in having dependent and independent variables. The dependent variables are the variables that we aim to predict, whereas the independent variables are the ones that enter the model and can help to explain the dependent variables.

The values of the dependent and independent variables were collected on a quarterly basis[1] to allow a greater amount of data available for the creation of the model, extracting all the information available at the time. It was possible to gather information from December of 2003 onwards, meaning that there were more than 15 years of historical data. The model is based on historical data from the Portuguese banking sector to avoid capturing unique effects from the past that, possibly, will not be felt in the same way at CGD under similar macroeconomic circumstances.

The model developed covers the projection of credit at CGD granted to individuals and non-financial companies, including credits to the Public Sector, and of deposits, both on time and on demand. In total, this projection covers about 80% of the balance sheet at Headquarters. In order to meet the necessary input data for the Strategy, Planning and Control Department (SPCD) P&L projection tool, the model allows to obtain the evolution of the portfolio volume of:

- Housing credit (in this case, the volume of new production is projected);
- Consumer credit;
- Other loans to individuals (excluding housing and consumption);
- Credits to large companies;
- Credit to small and medium companies;
- Credit to commercial real estate;
- Credit to the public sector;
- Credit to other companies;
- Term deposits;
- Demand deposits.

And the evolution of new production spread rates for the segments listed above (except demand deposits).

For the construction of the model, macroeconomic indicators related to the Portuguese market and the Eurozone are taken as independent variables and indicators of the country's banking sector are taken as dependent variables.

Regarding the historical data, most of it was collected from sources such as Bloomberg, Banco de Portugal, Eurostat and European Central Bank. Concerning future projections, values of macroeconomic projections in both scenarios for certain quarters, provided by Macroeconomic Studies Office (MSO) of CGD, are considered. Notice that the estimations were done for the three following years, so the projections of the variables had to be given for the same time periods.

[1]Whenever the source information was not on a quarterly basis, the value of the last day corresponding to each quarter was considered.

3.2.2.1 Dependent Variables

The dependent variables were modelled using data from the Portuguese banking system, collected exclusively from Banco de Portugal.

The models have as dependent variables, series referring to the portfolio volume and new production spread rates for each of the 10 segments defined above (with the exception of the spread rates for demand deposits), which are represented in Table 3.1.

Table 3.1: Dependent Variables

	Dependent Variable	Description
Volumes	**Vol_Consumer_Grow_YoY**	Annual growth rate of consumer credit (volume)
	Vol_Other_Grow_YoY	Annual growth rate of credit for other purposes (volume)
	Vol_SME_LC_Grow_YoY	Annual growth rate of credit to non-financial corporations (volume)
	Vol_Public_Sector_Grow_YoY	Annual growth rate of credit to general government except central government (volume)
	Vol_Real_State_Grow_YoY	Annual growth rate of credit to real estate activities (volume)
	New_Vol_Mortgage_Grow_YoY	Growth rate of new production of housing loans (volume)
	Vol_Term_Deposits_Grow_YoY	Annual growth rate of term deposits (volume)
	Vol_Sight_Deposits_Grow_YoY	Annual growth rate of sight deposits (volume)
Spreads	**Spread_New_Term_Dep_Grow_YoY**	Term Deposits of New Production Spread Rate
	Spread_New_Mortgage_Grow_YoY	Spread Rate of New Production of Home Loans
	Spread_New_Consumer_Grow_YoY	Spread Rate of New Consumer Credit Production
	Spread_New_Other_Grow_YoY	Spread Rate of New Credit Production for Other Purposes
	Spread_New_NFCup1M_Grow_YoY	Spread Rate of New Credit Production to Non-Financial Companies up to 1M€
	Spread_New_NFCover1M_Grow_YoY	Spread Rate of New Credit Production to Non-Financial Companies above 1M€

After collecting the information on a quarterly basis, similarly to what was done with the independent variables, it was necessary to determine the annual growth rate. For the volumes, it was necessary to apply the following formula:

$$volume_{t(tx)} = \frac{volume_t - volume_{t-12}}{volume_{t-12}} \qquad (target) \tag{3.1}$$

Where:

- $volume_{t(tx)}$, is the annual growth rate of the respective volume;
- $volume_t$, is the value of the given volume in month t;
- $volume_{t-12}$, is the value of the given volume existing in the 12th month prior to month t.

To determine spreads average rate[2], as the information on average spreads at the system level was not available, the spreads were obtained from the average value in the 12 months preceding month t of the interest rate by deducting the average of the relative Euribor 3M, Euribor 6M and Euribor 12M relative to the previous 12 months:

$$Spread_{nt} = avg(Interest\_rate_{n(t;t-12)}) - \frac{(avg(Euribor3M_{t;t-12}) + avg(Euribor6M_{t;t-12}) + avg(Euribor12M_{t;t-12})}{3} \tag{3.2}$$

[2]A spread rate is a component of the interest rate, set by the bank when granting a loan, that consists in the difference between the bid and purchase prices of a particular asset or financial instrument, in short, it corresponds to the financial institution's profit margin.

Where:

- $Spread_{nt}$, is the respective spread relative to month t;
- $avg(Interest\_rate_{n(t;t-12}$, is the average value of interest rates in the 12 months preceding month t (including month t);
- $avg(Euribor3M\_n(t;t-12$, is the average of the Euribor3M rates existing in the 12 months preceding month t (including month t);
- $avg(Euribor6M\_n(t;t-12$, is the average of the Euribor6M rates existing in the 12 months preceding month t (including month t);
- $avg(Euribor12M\_n(t;t-12$, is the average of the Euribor12M rates existing in the 12 months preceding month t (including month t);

The differences in spreads at 1 year (in relation to the same period of the previous year) result from the model's regressor:

$$Spread_{nt(1yeardifference)} = spread_{nt} - spread_{nt-12} \qquad (target) \tag{3.3}$$

Where:

- $Spread_{nt}$ is the respective spread relative to month t;
- $Spread_{nt-12}$, is the respective spread relative to the 12^{th} month prior to month t.

3.2.2.2 Independent Variables

Concerning the independent variables, the macroeconomic variables of the previous methodological version were used, as well as additional variables that proved necessary in the methodological test that took place during the revision of the model.

The macroeconomic variables were selected according to their economic meaning and were discussed internally between RMD and MSO in order to select the ones more relevant to the prediction of the dependent variables.

The set of the independent macroeconomic variables considered, as well as the exact sources, is detailed in Table 3.2:

Table 3.2: Independent Variables

Independent Variable	Description	Source
GDP_perc_yoy	Portuguese real GDP - 1 year growth	Bloomberg
Inflation_yoy	Portuguese inflation - 1 year growth	Bloomberg
Unemploy	Portuguese unemployment rate	Bloomberg
Euribor3M	Euribor 3 months	Bloomberg
Germany10Y	German Bond 10Y	Bloomberg
Portugal10Y	OT Portuguese	Bloomberg
PT_GER_10Y	OT-Bond Spread	computed as the difference between Portugal10Y and Germany10Y
resid_ppty_prices_EURO	Indices that measure the rate at which residential property prices change in the Eurozone	European Central Bank
racio_credito_vencido	Ratio of non-performing credit to total credit granted	Banco de Portugal
cmrcial_ppty_prices_EURO	Eurozone commercial property price index	European Central Bank
endividamento_familias	Families' inability to fulfil their obligations	Eurostat
rendimento_disponivel	Household disposable income	Eurostat
precos_residenciais	Residential Property Prices	Bloomberg
indice_prd_construcao	Construction production index	Eurostat
indice_prd_empresas	Production index in the industrial sector	Eurostat
CH_particulares_PT	Private mortgage loans in Portugal	Banco de Portugal
CH_particulares_EURO	Home loans in the Eurozone	Banco de Portugal
CC_particulares_PT	Consumer credit to individuals in Portugal	Banco de Portugal
CC_particulares_EURO	Consumer credit to individuals in the Eurozone	Banco de Portugal
TX_JR_habitacao_PT	Interest rate on home loans in Portugal	Banco de Portugal
TX_JR_habitacao_EURO	Interest rate on home loans in the Eurozone	Banco de Portugal
TX_JR_consumo_PT	Interest rate on consumer loans in Portugal	Banco de Portugal
TX_JR_consumo_EURO	Interest rate on consumer loans in the Eurozone	Banco de Portugal

The variables presented are the originals collected from the source. Yet, if the variable was already in 1-year growth, only the annual lag was added, otherwise, it was necessary to create the annual growth rate and the 1-year lag of the annual growth rate. The annual variation, that is, the annual growth rate is calculated by:

$$AnnualGrowthRate = \frac{(EndingValue - BegginingValue)}{BeginningValue} \tag{3.4}$$

The annual lag, consists on the value of the specific variable for the same period in the previous year.

3.2.2.3 Final Input Data Set

All available historical information for both dependent and independent variables was then compiled into an Excel file, which was later imported to SAS, creating a SAS table with the input data. The historical information was available from the last quarter of 2003 until the second quarter of 2020, although a few variables only had information since 2005. Notice that at least 30 observations were needed, since this is the limit for statistical analysis and results to be significant and reliable.

After analysing the quality of the information collected, some outliers were found. Since, the inclusion of these observations would have a considerable impact in the model and potentially in the statistical significance of the results, an outlier's analysis was performed and these points were removed from the historical information in order to control the variability in the

data and prevent distortions in statistical analysis and consequent reduction of the statistical power.

Then, in order to model the historical data, it was necessary to create a unique input data set which contained all the variables needed already cleaned, that is, without any outliers. At the time of extraction, it was found that some of the variables were only available until the second quarter of 2020, so the historical information considered was limited to June of 2020. However, the last 8 observations (from September of 2018 until June of 2020) were included in the out-of-time sample. The remaining quarters of 2020 as well as the years of 2021, 2022 and 2023 were the periods which were intended to be estimated. Overall, a worksheet in Excel was created, having all quarters from December of 2003 until December of 2023 (which was the last estimation period). The input data set had around 70 variables, including dependent variables, original and transformed independent variables.

The following figure illustrates the executed procedures to achieve the final input data set to apply time series forecasting.

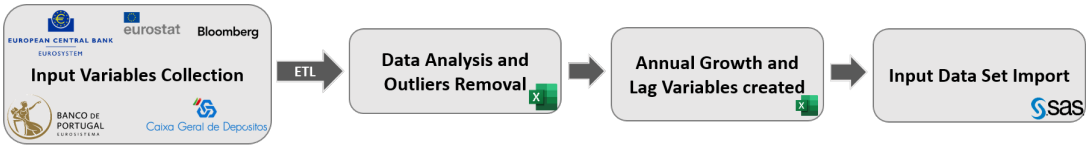

Figure 3.2: Process to get Final Data Set

3.2.3 Data Description

As mentioned in section 3.2.2, the necessary data set for the application of time series to forecast the target credit items included dependent and independent variables. The dependent variables consist in the targets that are aimed to be predicted, which are the CGD's credit volumes and spreads. The independent variables are the exogenous macroeconomic variables that can help to forecast the targets, such as unemployment rate, GDP, inflation, among others. It was necessary to collect the historical information on a quarterly basis, and when this was not possible, with only monthly data available, the last day of the month of each quarter was considered.

The main challenge relied on finding historical data in the pretended basis, but more importantly, to find independent variables that could be useful to explain the respective targets. The independent variables had to be highly correlated with the dependent variables, and with the correct economic rationale. Then, an outlier analysis was made, and some observations were removed since it was concluded that these could impact the obtained results. These were the main procedures done to reach the input data set.

The specific historical data used, as well as the procedures made to obtain the input data set for the modelling process, and the sources used for the collection of the information are also described in section 3.2.2.

3.2.4 Practical Example

In order to describe the procedures executed to achieve a final time series model, an example for the target variable of the new production of housing loans growth rate ("New_Vol_Mortgage") will be described.

Firstly, it was essential to understand if the dependent variable was stationary, and if not, differentiate it until it was. For this purpose, the graphical observation, the Dickey-Fuller Test and the Autocorrelation Function of the new production of housing loans ("New_Vol_Mortgage") were taken into consideration. Additionally, it was indispensable to analyse the series observations graph in order to verify if there was any occurrence of seasonality. The graphical observation of the original series is represented in Figure 3.3:

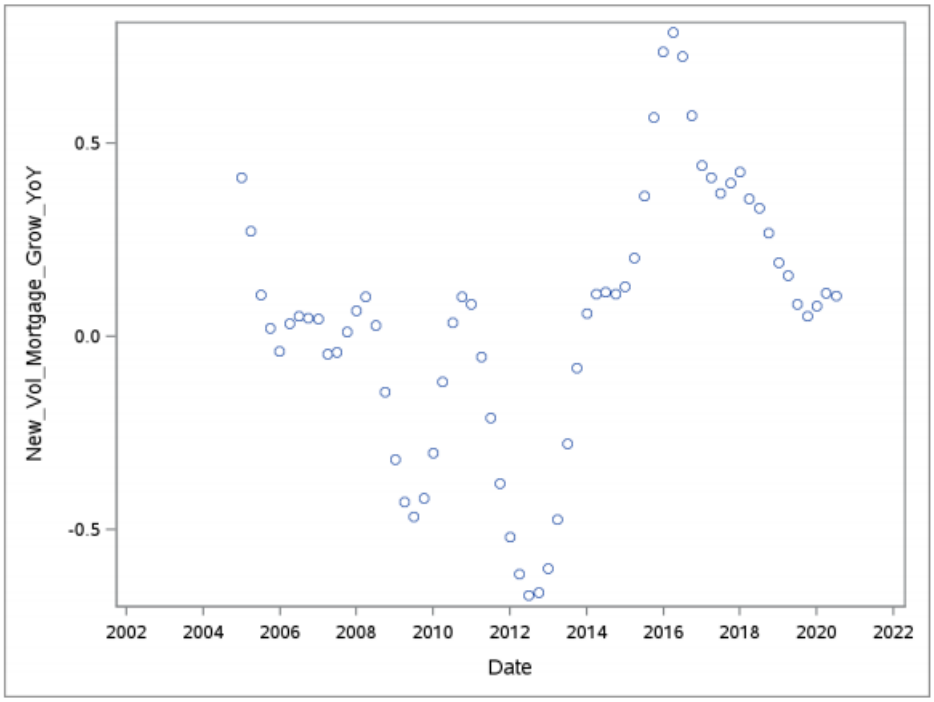

Figure 3.3: Original Series of New Production of Housing Loans ("New_Vol_Mortgage")

Even though it is possible to conclude just by the visualization that the series is not stationary, the Dickey-Fuller Test and the ACF were also analysed, obtaining the results presented in Figure 3.4:

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-3.3686	0.2037	-1.46	0.1328		
	1	-25.1430	0.0001	-3.60	0.0005		
Single Mean	0	-3.3088	0.6099	-1.41	0.5701	1.07	0.7988
	1	-25.9752	0.0011	-3.63	0.0078	6.58	0.0035
Trend	0	-4.7493	0.8293	-1.91	0.6364	2.28	0.7230
	1	-29.0229	0.0042	-3.91	0.0174	7.66	0.0209

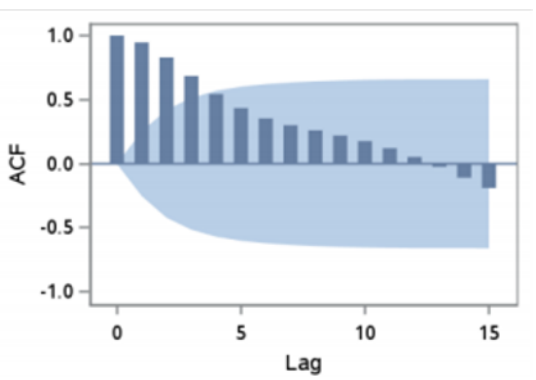

Figure 3.4: Dickey-Fuller and ACF for the original series of New Production of Housing Loans ("New_Vol_Mortgage")

As it is evidenced, the series is not stationary due to the facts:

- *p-value* of Tau for the Zero Mean is higher than 0.05, revealing the acceptance of the null hypothesis of Dickey-Fuller;
- ACF decreases very slowly, between lags.

Considering that the non-stationarity of the time series was proven, it was necessary to differentiate it. The first, the second and the third differences were applied, obtaining the results represented in Figures 3.5, 3.6 and 3.7, respectively:

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-12.8500	0.0103	-2.77	0.0063		
	1	-44.4364	<.0001	-4.88	<.0001		
Single Mean	0	-12.8178	0.0555	-2.74	0.0736	3.79	0.1188
	1	-44.4107	0.0006	-4.83	0.0002	11.69	0.0010
Trend	0	-12.7226	0.2407	-2.67	0.2506	3.70	0.4465
	1	-44.3128	0.0001	-4.76	0.0015	11.47	0.0010

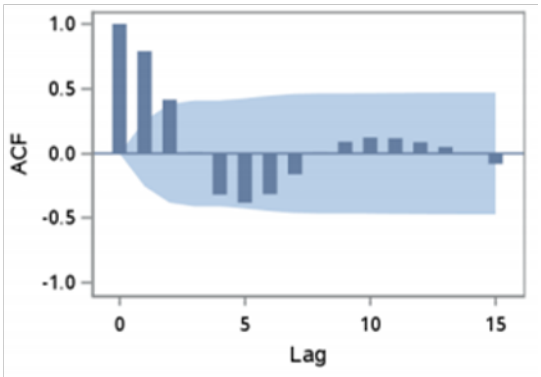

Figure 3.5: Dickey-Fuller and ACF for the first-differenced series of New Production of Housing Loans ("New_Vol_Mortgage")

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-35.5833	<.0001	-4.98	<.0001		
	1	-37.8708	<.0001	-4.32	<.0001		
Single Mean	0	-35.6446	0.0006	-4.94	0.0002	12.20	0.0010
	1	-37.8697	0.0005	-4.28	0.0011	9.17	0.0010
Trend	0	-35.7995	0.0005	-4.92	0.0009	12.16	0.0010
	1	-37.9784	0.0002	-4.24	0.0072	9.00	0.0010

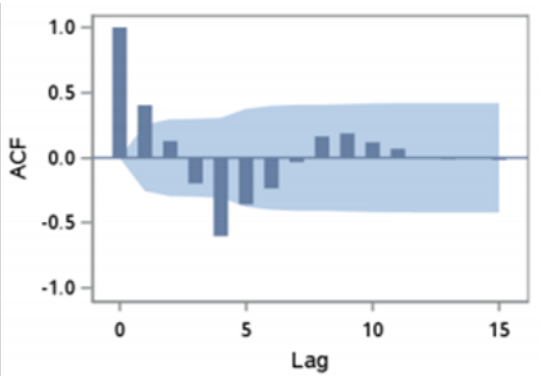

Figure 3.6: Dickey-Fuller and ACF for the second-differenced series of New Production of Housing Loans ("New_Vol_Mortgage")

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	0	-74.1741	<.0001	-10.08	<.0001		
	1	-75.5744	<.0001	-6.00	<.0001		
Single Mean	0	-74.1615	0.0005	-10.00	0.0001	50.04	0.0010
	1	-75.5458	0.0005	-5.95	0.0001	17.69	0.0010
Trend	0	-74.1533	0.0001	-9.91	<.0001	49.13	0.0010
	1	-75.5543	0.0001	-5.89	<.0001	17.35	0.0010

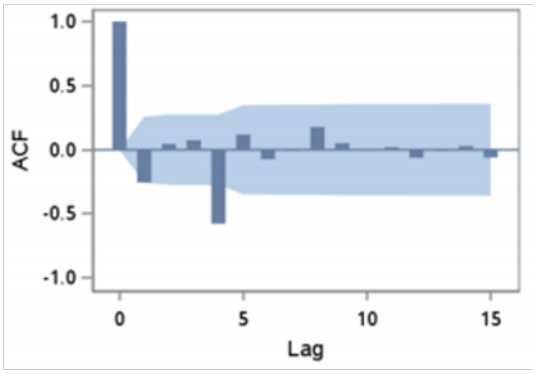

Figure 3.7: Dickey-Fuller and ACF for the third-differenced series of New Production of Housing Loans ("New_Vol_Mortgage")

After analysing the results, it was decided to start testing with the application of the third difference, since it was the one that left no doubt on the presence of stationarity in the series, according to the Dickey-Fuller test and to the ACF. Then, the independent variables used in the model also had to be differentiated three times, and its correlations were analysed.

Effectively, there was no evidence of seasonality in the graphical visualization of the annual growth rate of the new production of housing loans ("New_Vol_Mortgage") that justified the application of SARIMAX, so ARIMAX was the chosen model to put in practice since it allows the inclusion of independent variables, which enriches the obtained predictions. Therefore, the Next step was to select the exogenous variables to include. The selection started by testing the variables that resulted from the application of the Stepwise Method, checking if they had a considerable correlation with the target and with the correct economic rationale. For this particular case, the Stepwise results obtained are presented in Table 3.3:

Table 3.3: Stepwise Results for the third-differenced series of New Production of Housing Loans ("New_Vol_Mortgage")

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	_3dif_TX_JR_consm_EU		1	1	9.5682		0.0020
2	_3dif_resid_ppty_pri		1	2	4.8232		0.0281
3	_3dif_Euribor3M_Var_		1	3	6.0645		0.0138
4	_3dif_indice_prd_con		1	4	4.4985		0.0339
5		_3dif_TX_JR_consm_EU	1	3		1.7485	0.1861
6	_3dif_PT_GER_10Y_Var		1	4	7.1265		0.0076
7	_3dif_CH_parti_EUR_V		1	5	4.4175		0.0356
8	_3dif_TX_JR_consm_PT		1	6	4.7912		0.0286
9		_3dif_indice_prd_con	1	5		2.9910	0.0837
10	_3dif_inflation_yoy		1	6	6.7352		0.0095
11	_3dif_racio_crd_venc		1	7	5.4382		0.0197
12	_3dif_TX_JR_habit_PT		1	8	3.2340		0.0721
13		_3dif_TX_JR_habit_PT	1	7		3.2417	0.0718

For instance, the only variables that the last iteration of this method concludes as being useful for the model are the ones which were not removed, namely:

- **_3dif_resid_ppty_pri** → 3rd difference of Indices that measure the rate at which residential property prices change in the Eurozone – annual growth rate;
- **_3dif_Euribor3M_Var_** → 3rd difference of Euribor 3 months - annual growth rate lagged 1 year;
- **_3dif_PT_GER_10Y_Var** → 3rd difference of OT-Bond Spread - annual growth rate lagged 1 year;
- **_3dif_CH_parti_EUR_V** → 3rd difference of Home loans in the Eurozone - annual growth rate lagged 1 year;
- **_3dif_TX_JR_consm_PT** → 3rd difference of Interest rate on consumer loans in Portugal - annual growth rate lagged 1 year;
- **_3dif_inflation_yoy** → 3rd difference of Portuguese inflation - annual growth rate;
- **_3dif_racio_crd_venc** → 3rd difference of Ratio of non-performing credit to total credit granted - annual growth rate lagged 1 year.

Afterwards, the correlations of each of these variables were verified to select the ones which respected the requirements of having a considerable correlation percentage with the target, the correct economic rationale and low correlation between them. The correlations obtained are displayed in the last column of Table 3.4:

`\hline`:

Table 3.4: Stepwise SAS Results

Variable Entered in Stepwise	Variable Removed from Stepwise	Correlation with target
_3dif_TX_JR_consm_EUR_Var1Y (Interest rate on consumer loans in the Eurozone – annual growth rate – 3rd difference)		-25,63%
_3dif_resid_ppty_prices_EURO (Indices that measure the rate at which residential property prices change - Eurozone – 3rd difference)		-24,67%
_3dif_Euribor3M_Var_Lag1Y (Euribor 3 months – annual growth rate – 1 year lag – 3rd difference)		3,28%
_3dif_indice_prd_const_Var_Lag1Y (Construction production index – annual growth rate – 1 year lag – 3rd difference)		-29,39%
	_3dif_TX_JR_consm_EUR_Var1Y (Interest rate on consumer loans in the Eurozone – annual growth rate – 3rd difference)	-25,63%
_3dif_PT_GER_10Y_Var_Lag1Y (OT-Bond Spread – annual growth rate – 1 year lag – 3rd difference)		6,17%
_3dif_CH_parti_EUR_Var_Lag1Y (Home loans in the Eurozone – annual growth rate – 1 year lag – 3rd difference)		11,25%
_3dif_TX_JR_consm_PT_Var_Lag1Y (Interest rate on consumer loans in Portugal – annual growth rate – 1 year lag – 3rd difference)		3,61%
	_3dif_indice_prd_const_Var_Lag1Y (Construction production index – annual growth rate – 1 year lag – 3rd difference)	-29,39%
_3dif_inflation_yoy (Portuguese inflation – annual growth rate – 3rd difference)		37,08%
_3dif_racio_crd_vencid_Var_Lag1Y (Ratio of non-performing credit to total credit granted – annual growth rate – 1 year lag – 3rd difference)		35,60%
_3dif_TX_JR_habit_PT_Var1Y (Interest rate on home loans in Portugal – annual growth rate – 3rd difference)		4,67%
	_3dif_TX_JR_habit_PT_Var1Y (Interest rate on home loans in Portugal – annual growth rate – 3rd difference)	4,67%

The rows coloured in dark grey are representing the correlations that were removed by the Stepwise procedure, the variables from the white rows were not considered since the correlation was significantly low, and the light grey rows are the variables that had at least a threshold of 20% correlation with the dependent variable. Initially it was intended to consider percentages higher or lower than 50% and -50% respectively, but since it was extremely

difficult to achieve these values, the threshold was reduced to 20%.

In this case, the three variables coloured in light grey were tested in the ARIMAX procedure. Nevertheless, when carefully analysing the outcomes, it was realised that the only variable that had a correct economic rationale (coherent sign of the coefficient) with the new production of housing loans ("New_Vol_Mortgage"), was the annual growth of the inflation rate. For instance, it is not economically logical that the residential property prices in the Eurozone ("3dif_resid_ppty_pri") is inversely proportional to the new production of housing loans ("New_Vol_Mortgage") , it should be the opposite, and it does not make sense in economic terms that the ratio of non-performing credit to total credit granted growth rate ("_3dif_racio_crd_venc") values are directly proportional to the values of new production of housing loans ("New_Vol_Mortgage"). Therefore, although the model passed in all the statistical tests, the theoretical rationale was not consistent given the economic context and projections.

The next step was testing several models with various sets of independent variables taking into consideration the ones that had the correct economic rationale given the economic context. Initially, this approach that did not use the Stepwise Method, was tested using also the third difference. Nonetheless, the results were not as good as expected and by analysing the initial autocorrelation output for the second difference, which is usually considered to be enough to differentiate a series [Bleikh and Young, 2016], more alternatives were tried using the second difference.

Several attempts were also tested using the series differentiated twice, and again, the next step was defining the exogenous variables to include. With the purpose of choosing the independent variables, the correlations were computed, and the ones that surpassed the threshold of 20% are represented in Table 3.5:

Table 3.5: Independent Variables with correlations higher than 20% with target

Independent Variable	Correlation with target
_2dif_GDP_perc_yoy (Portuguese real GDP – annual growth rate – 2nd difference)	20,38%
_2dif_inflation_yoy (Portuguese inflation – annual growth rate – 2nd difference)	29,09%
_2dif_Inflation_yoy_Lag1Y (Portuguese inflation – annual growth rate –1 year lag – 2nd difference)	-26,52%
_2dif_Unemploy_Var_1Y (Portuguese unemployment rate – annual growth rate – 2nd difference)	-30,89%
_2dif_Unemploy_Var_Lag1Y (Portuguese unemployment rate – annual growth rate – 1 year lag – 2nd difference)	35,69%
_2dif_Euribor3M_Var_1Y (Euribor 3 months – annual growth rate – 2nd difference)	20,58%
_2dif_racio_crd_vencid_Var_1Y (Ratio of non-performing credit to total credit granted – annual growth rate – 2nd difference)	-21,21%
_2dif_racio_crd_vencid_Var_Lag1Y (Ratio of non-performing credit to total credit granted – annual growth rate – 1 year lag – 2nd difference)	23,66%
_2dif_endivid_familias (Families' inability to fulfil their obligations – annual growth rate – 2nd difference)	31,75%
_2dif_Endivid_familia_Var_Lag1Y (Families' inability to fulfil their obligations – 2nd difference)	-35,49%
_2dif_indice_prd_const_Var_Lag1Y (Construction production index – annual growth rate – 1 year lag – 2nd difference)	-21,45%
_2dif_TX_JR_habit_PT_Var1Y (Interest rate on home loans in Portugal – annual growth rate – 2nd difference)	23,11%
_2dif_TX_JR_habit_PT_Var_Lag1Y (Interest rate on home loans in Portugal – annual growth rate – 1 year lag – 2nd difference)	-44,55%
_2dif_TX_JR_habit_EUR_Var_Lag1Y (Interest rate on home loans in the Eurozone – annual growth rate – 1 year lag – 2nd difference)	-30,66%

After carefully analysing the variables and the corresponding correlation percentages between the target and the independent variables, the correlations between the independent variables were also examined. The goal was to use independent variables which were not correlated with each other, in order to avoid problems when fitting the model since multicollinearity influences the regression coefficients' estimations. Taking this into consideration, many experiments were done, yet, the following independent variables were the ones used in the final model:

- **_2dif_endivid_familias** → Families' inability to fulfil their obligations – annual growth rate – 2nd difference;
- **_2dif_racio_crd_vencid_Var_1Y** → Ratio of non-performing credit to total credit granted – annual growth rate – 2nd difference.

The correlations between the independent variables used were verified, using SAS by executing VIF Multicollinearity Test and reaching the output represented in Table 3.6:

Table 3.6: VIF Multicollinearity Test Results for the second-differenced series of the Independent Variables

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	0.00166	0.00782	0.21	0.8327	.	0
_2dif_racio_crd_vencid_Var_1Y	_2dif_racio_crd_vencid_Var_1Y	1	-0.83687	0.59167	-1.41	0.1626	0.98387	1.01640
_2dif_endivid_familias	_2dif_endivid_familias	1	3.22640	1.34813	2.39	0.0200	0.98387	1.01640

The value of the Variance Inflation should not be greater than 10, which is respected since the value obtained was 1.01640, concluding that both variables can be used simultaneously without compromising the estimation of the parameters of the model.

Then, it was necessary to determine the p (AR) and q (MA) orders. Initially, the approach used was analysing the autocorrelation function and the partial correlation function. However, later on the project, a more effective and automatic technique was found and implemented. As referred in section 2.3.1.2.4, the methods executed automatically in SAS to determine orders p (AR) and q (MA) were:

- **MINIC** - Minimum Information Criterion Method;
- **ESACF** - Extended Sample Autocorrelation Function Method;
- **SCAN** - Smallest Canonical Correlation Method.

The output obtained after executing MINIC, ESACF and SCAN in SAS, is illustrated in Table 3.7:

Table 3.7: MINIC, SCANF and ESACF Results

ESACF Probability Values						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	0.0015	0.4012	0.1629	0.0005	0.0620	0.2754
AR 1	0.1401	0.3696	0.4022	0.0001	0.8516	0.2834
AR 2	0.1710	0.6447	0.5681	<.0001	0.4426	0.5247
AR 3	0.0032	0.3309	0.0831	0.0004	0.3071	0.3829
AR 4	0.0114	0.9139	0.1727	0.0018	0.1282	0.0906
AR 5	0.0019	0.7518	0.7966	0.0397	0.7865	0.1128

Squared Canonical Correlation Estimates						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	0.1928	0.0200	0.0569	0.4173	0.1776	0.0718
AR 1	0.0083	0.1249	0.0997	0.4217	0.0010	0.0511
AR 2	0.1213	0.0058	0.1225	0.3587	0.0563	0.0253
AR 3	0.2872	0.2352	0.1770	0.2536	0.0045	0.0090
AR 4	0.0511	0.0006	0.0254	0.1160	0.0061	0.0105
AR 5	0.0207	0.0287	0.0018	0.0743	0.0027	0.0126

SCAN Chi-Square[1] Probability Values						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	0.0008	0.3872	0.1497	<.0001	0.0504	0.2645
AR 1	0.5099	0.0103	0.0239	<.0001	0.8558	0.2171
AR 2	0.0102	0.6399	0.0225	0.0001	0.1587	0.3734
AR 3	<.0001	0.0016	0.0071	0.0025	0.6967	0.5756
AR 4	0.1089	0.8816	0.3389	0.0221	0.6400	0.6069
AR 5	0.3165	0.3068	0.7854	0.0917	0.7567	0.5755

Minimum Information Criterion						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	-5.63507	-5.67423	-5.60009	-5.57965	-6.00311	-6.06626
AR 1	-5.9898	-5.93212	-5.89174	-5.96737	-6.32124	-6.26728
AR 2	-6.04695	-6.08792	-6.0485	-6.09767	-6.28726	-6.32614
AR 3	-6.11996	-6.17552	-6.10474	-6.03005	-6.25663	-6.2896
AR 4	-6.37097	-6.29744	-6.24376	-6.19773	-6.21697	-6.21602
AR 5	-6.34505	-6.28795	-6.28842	-6.2137	-6.18078	-6.15845

Extended Sample Autocorrelation Function						
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR 0	0.4364	0.1355	-0.2282	-0.5895	-0.3816	-0.2373
AR 1	0.2046	0.1714	-0.1180	-0.5356	-0.0327	-0.2366
AR 2	-0.1917	0.0826	0.0865	-0.5700	-0.1372	-0.1425
AR 3	-0.4165	0.1679	0.2789	-0.5224	0.2152	-0.2026
AR 4	0.3614	-0.0174	-0.2152	-0.4480	-0.2879	-0.3029
AR 5	0.4492	-0.0497	-0.0379	-0.3058	0.0475	-0.2730

ARMA(p+d,q) Tentative Order Selection Tests					
SCAN			ESACF		
p+d	q	BIC	p+d	q	BIC
0	4	-6.00311	0	4	-6.00311
5	0	-6.34505	1	4	-6.32124
			2	4	-6.28726
			3	4	-6.25663
			4	4	-6.21697
			5	4	-6.18078

By observing the output of the last table, the various possible combinations of orders p and q by SCAN and ESACF are represented. Regarding MINIC method, the orders suggested correspond to the lowest BIC value, which, in this case, are AR(0) and MA(1). Even though these were the possible combinations tested, for confidential reasons, the p and q orders used in this final model will not be identified. With the aim of choosing the best orders of p and q, that is, after executing the whole process for the several attempts, it was necessary to analyse all the residuals tests, which are subsequently described specifically for the final model.

Concerning Ljung-Box test, whose output is displayed in Table 3.8, the conclusion after validating that the *p-value* is greater than 0.05, was that the residuals follow a white noise process and therefore the model does not show fit failure.

Table 3.8: Ljung-Box Results

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	9.38	5	0.0947	0.288	0.044	-0.022	-0.171	-0.155	-0.150
12	17.85	11	0.0851	-0.206	-0.198	0.123	0.052	0.096	0.141
18	19.39	17	0.3069	0.049	0.052	-0.087	-0.054	-0.013	-0.062
24	30.81	23	0.1275	-0.246	-0.081	0.109	0.117	0.180	-0.039

By the results obtained from DW, it was verified that the residuals are independent meaning that they do not have autocorrelation. This conclusion can be confirmed by checking if the *p-value* is greater than 0.05 in Table 3.9:

Table 3.9: DW Results

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	1.9538	0.4329	0.5671
2	1.9651	0.5059	0.4941
3	2.1401	0.7878	0.2122
4	1.8545	0.4579	0.5421

Concerning Shapiro-Wilk test, it was intended to have a value of W greater than 0.947 to validate that the residuals follow a normal distribution, as mentioned in section 2.3.1.3.3. This condition was confirmed after obtaining the following output, in Table 3.10, where W has a value of 0.98.

Table 3.10: Shapiro-Wilk Test Results

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.982539	Pr < W	0.6267
Kolmogorov-Smirnov	D	0.092713	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.062893	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.339529	Pr > A-Sq	>0.2500

The last residuals test executed, was the ARCH test whose purpose is to check if residuals are heteroscedastic or homoscedastic. Since the outcomes from SAS, displayed below, in Table 3.11, demonstrate the *p-value* for all orders was greater than 0.05, it was confirmed the presence of homoscedasticity in the residuals.

Table 3.11: ARCH Test Results

Tests for ARCH Disturbances Based on OLS Residuals				
Order	**Q**	**Pr > Q**	**LM**	**Pr > LM**
1	0.2891	0.5908	0.3369	0.5616
2	0.3535	0.8380	0.4409	0.8022
3	0.7866	0.8527	0.6795	0.8780
4	2.1497	0.7083	2.6188	0.6235
5	2.1601	0.8266	2.7835	0.7333
6	2.2638	0.8939	3.4852	0.7459
7	3.1459	0.8712	5.2900	0.6246
8	3.3154	0.9130	5.2919	0.7260
9	3.3232	0.9501	5.9623	0.7437
10	3.3241	0.9727	6.3505	0.7850
11	5.7240	0.8911	6.8751	0.8091
12	5.9796	0.9171	8.2325	0.7667

Consequently, it was concluded that the model under study passed all the residuals statistical tests, indicating it had a considerably good fit. Then, the performance indicators were evaluated to determine the forecasting power of the model. The output presented in Table 3.12, shows the AIC indicator, which is very negative as pretended, since the lower the AIC the better the model. Actually, this performance indicator can be used to compare several models in order to choose the best one.

Table 3.12: AIC Results

Constant Estimate	0.000795
Variance Estimate	0.002177
Std Error Estimate	0.046653
AIC	-169.736
SBC	-159.885
Number of Residuals	53

Another performance indicator that was analysed, is the MSE which was manually computed with the aim of comparing the several models obtained for each target. The value of MSE was of 0.001971 when using all the historical observations and 0.003453 when the out-of-time sample was included. As it was explained previously, the best model is the one with a value closer to zero, as it happens in this model, proving it has a reliable fit.

Furthermore, all these outcomes were detailed in an Excel file together with the graphical visualizations of the historical observations, the predictions on a baseline and adverse scenario, and the previous estimations obtained with multiple linear regression (the previous

method of estimation). The resume of all these analysis regarding the model for the target of new production of housing loans ("New_Vol_Mortgage") are represented in Figure 3.8[3].

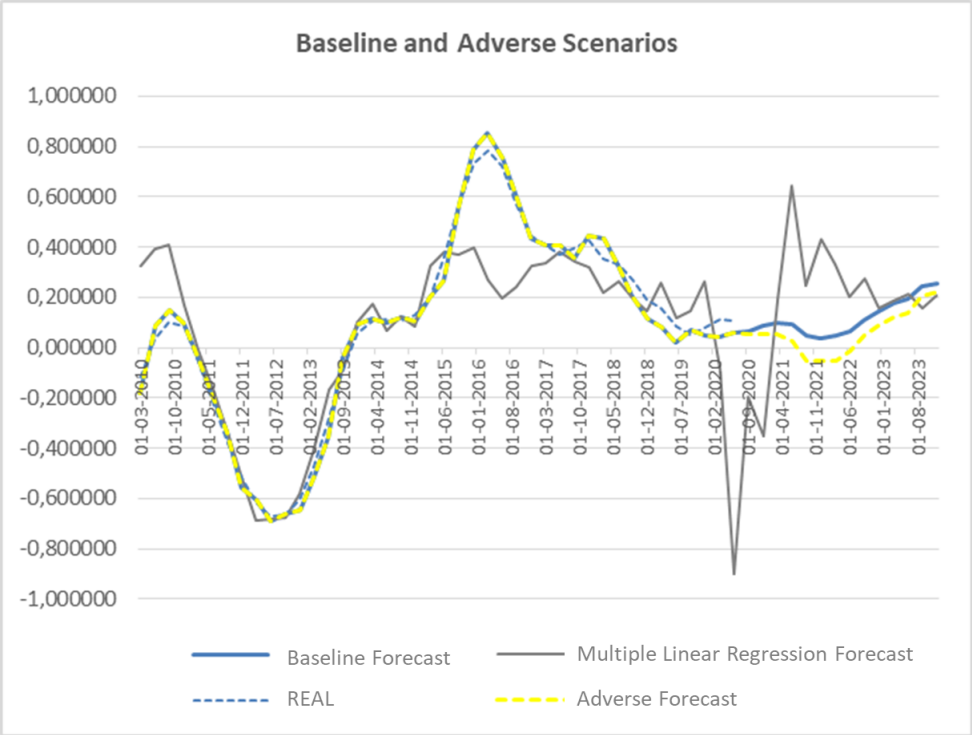

Figure 3.8: Forecasting vs Real and Previous Methodology Results

As shown in Figure 3.8, the baseline scenario has higher values than the adverse scenario meaning that in a more favourable economic context the new production of housing loans is higher than in a worse economic scenario. Additionally, it is evident that time series forecasting results overcame the predictions obtained with multiple linear regressions, proving it is a more reliable and robust method, being a considerable improvement in the estimation of the main credit items.

3.2.5 Results and Discussion

Among the time series models described in section 2.3.1.1, the ARIMAX models were the chosen ones to forecast the target variables. This option was based on the need to obtain models that allowed the inclusion of exogenous variables to which the baseline and adverse scenarios could then be applied to obtain projections for the following three years. Additionally, the ARIMAX models were those that reached better results in the scope of the methodological simulation that took place during the development of the new time series methodology.

[3]The multiple linear regression forecast, illustrated in Figure 3.8, is represented by the equation:

$$New\_Vol\_Mortgage = 0,051 + 7,166 \times GDP\_perc\_yoy - 10,740 \times Inflation\_yoy\_1Y\_Lag1Y \quad (3.5)$$

Since the graphical observations of the series did not evidence the presence of seasonality and that the SARIMAX models did not show good results in the methodological simulation, combined with the fact that the dependent variables represent year-on-year variations, thus removing out any seasonality effects that might be present, and the fact that there is lack of granularity in the data (quarterly data instead of daily or weekly, where the existence of seasonality is more likely), this type of models was not considered to calculate the forecasts of the target variables.

In order to determine the autoregressive (AR) and moving average (MA) orders of the models, in an early stage the significant correlations of the autocorrelation and partial autocorrelation functions were analysed. However, at a later stage of the process, the more automated approach of the SCAN, ESACF and MINIC methods was applied in SAS.

Notably, with the aim of following EBA regulatory guidelines, the data sample was divided into a development sample and an out-of-time sample, in order to allow an analysis of forecast errors outside the sample considered for the development of the model. Thus, the estimation of the models was based on the development sample and the out-of-time period corresponded to the last 8 available historical observations. The prediction errors, in particular the MSE results were analysed for both samples.

Concerning the approach used to determine the independent variables to help predict the target variables, both Stepwise and manual selection methods were executed. Although Stepwise was applied to all target variables, it was found that in some situations, no statistically significant variable that could be considered in the models was returned. For these cases, the analysed model options were based only on the manual selection method.

It should be noted that the application of the Stepwise method took into account the differentiation of the series. That is, if for a certain target variable the application of the first difference was concluded, then the exogenous variables considered in the procedure were also differentiated once. Additionally, it was decided to introduce the entire available historical period of the variables in the stepwise procedure (instead of considering only the period corresponding to the development sample), thus taking advantage of all the available information (since there are not many observations) and considering the relative instability that the dependent variables' series presented. For the exogenous variables returned by the Stepwise, the correlations with the respective dependent variable were also taken into account, and those with correlations above 20% with the target variable were tested in the models.

Regarding the procedure that does not use Stepwise, it intends to initialize with the independent variables that present a high correlation with the target variable, confirming then whether the selected variables are statistically significant. Again, if they were not, they were removed until all of them were relevant for the model. Additionally, the correlations of the independent variables with each other were also taken into account, as well as whether the respective theoretical sense was economically coherent to predict the evolution of the dependent variable. It should be noted that the correlations were analysed taking into consideration the previously determined difference to be used for each series. That is, if for a certain target the use of the first difference was concluded, the correlations were analysed for all the differentiated variables once. Additionally, it was found that, as the series are

differentiated again and again, the correlations (with the target and between the independent variables) generally decrease. Thus, taking into account that the correlations are not generally high, it was necessary to consider correlations with the dependent variables below 50%.

This process involved several trials and adjustments in order to determine the best variables for the models. Initially, variables whose correlation with the target was greater than 20% and whose correlation with each other was less than 50% were considered. However, attempts were also made with exogenous variables with correlations lower than 20% as long as a logical theoretical sense was confirmed, providing a robust economic rationale to the models. For each tested model, it was taken into account whether the AR and MA estimators were significant as well as whether the exogenous variables were significant. Whenever these two criteria were not met, it was necessary to adjust the estimators and/or exogenous variables until they were all significant. To fulfil these criteria, a *p-value* of 0.2 was considered, since if it was considered a lower *p-value*, the available hypotheses would be very limited.

Once it was assured that both estimators and variables were significant, the VIF test was analysed to ensure the non-existence of multicollinearity in the model, in cases where more than one independent variable were considered. Then, the results of the tests for the residuals listed in section 2.3.1.3 were evaluated, namely if they followed a white noise process (Ljung-Box test), if they were independent (DW test), if they followed a normal distribution (Shapiro-Wilk test) and finally, if they were homoscedastic (White/ARCH test)[4]. In fact, it was not possible to guarantee that the models would pass all the listed tests since there was a reduced number of observations. However, an effort was made to ensure that the final models chosen passed as many tests as possible, not forgetting other factors such as the explanatory variables being functionally adequate to explain the target.

To assess the performance of the models, the performance indicators described in section 2.3.1.4 were analysed, namely, the AIC and the MSE criteria, seeking to ensure that the final models were those with the best indicators (with priority for the AIC) together with other criteria of expert judgment (namely, the economic meaning of the variables).

The choice of the final models to predict the target variables resulted from the combination of several criteria, namely:

- Present a theoretical rationale consistent with the economic context of forecasting the target variable;
- Present explanatory variables that are functionally adequate to explain the target;
- Allow for a behavioural distinction between the baseline and the adverse scenario (although it depends on the macroeconomic projections provided by MSO);
- Present acceptable statistical results;
- Present acceptable performance (with priority given to the AIC criteria).

[4]These tests are referenced in EBA guidelines, and since the good practices of Stress Test were taken into consideration, it was aimed that the achieved models respected these statistical tests.

In conclusion, the final models chosen for each target variable result from the best possible combination of all the criteria mentioned above. For this reason, the finals models selected for each target, which were always determined with the model owner, were not always the ones with the best statistical results and/or the best AIC, but the ones that globally represented the best combination of all criteria. However, an effort was made to certify that the models were statistically robust and ensured a good performance.

3.2.6 Conclusions

To sum up, the main purpose of this project was to implement a new methodology to predict the main balance sheet credit items (volumes and spreads) of CGD. Whereas before these were projected using multiple linear regressions, the enhancement consisted in applying time series forecasting. After verifying that this approach would provide significantly better results, it was implemented for CGD, and later on for some of the entities of the Group.

In summary, the work carried out intended to implement improvements in the results obtained for the following:

- Methodology more consistent with the nature of the data;
- Results summary document in articulation with estimation results, minimizing operational errors and synthesizing the results and selection of models for the estimation;
- Comparison of different estimates;
- Results with a better fit facing the historical data, when compared to the projections obtained with linear regressions;
- Collection of new macroeconomic variables that could help to predict more efficiently the targets;
- Projection of quarterly estimations for baseline and adverse scenarios, instead of annual projections.

The developed methodology employing time series was not only used to estimate the credit volumes and spreads of CGD and other entities, but it was also used for the purposes of CGD's Recovery Plan, reinforcing the fact that it was a robust approach to put in practice for several purposes.

4. Conclusions

This report aimed to describe the work developed, as well as the challenges encountered along the internship performed in the MDU at the RMD of CGD. In particular, the focus of this report is mainly on the project which purpose was to estimate the values of credit volumes and spreads of the main segments of CGD's balance sheet. There were three main goals established at the beginning of the internship. The first consisted in developing a forecasting model to be used in the ICAAP exercise.

Secondly, getting to know the different areas of the bank, in particular, find out what were their responsibilities and objectives, and which strategies they apply to achieve them. Finally, integrating the labour market, learning how to manage time effectively, defining a working plan evaluating impacts and solutions for continuous improvement.

The forecasting model developed was the Balance Sheet projection of credit items which was the main project described throughout this document. It turned out to be a challenging project which required a great sense of responsibility and commitment due to all the complexity and difficulties that arose during its development. The estimation of the credit volumes and spreads of CGD is extremely relevant for the bank's management and decision-making, supporting the strategical plan definition and budget.

Regarding the second objective, there were several online sessions where a specific area of the bank would be presented as well as their daily work and how it impacted the institution's business. Also, during the internship, there were various situations in which it was necessary to get in touch with other areas. For instance, in the Balance Sheet Projection project, it was essential to work with MSO since they were responsible to provide the macroeconomic scenarios evolution which were an input for the models. It was also necessary to work with MVO since they were responsible to review and validate the models developed.

Therefore, although there are many areas inside CGD, being impossible to get to know in detail all of them in a short period of time, it was possible to get a feeling of how other areas and the institution organization work.

Concerning the integration in the labour market, it was fairly easy to integrate the professional environment in the MDU of CGD due to its remarkable team spirit and working experience. Although there was a substantial self-improvement in establishing working plans and managing time in a more effective way, there are still certain skills that need to be improved. For instance, being able to prioritize tasks that are requested simultaneously without feeling excessively under pressure and better understand the concepts dealt with in detail in each project, especially regarding more specific topics in the banking area, are two aspects which need to be enhanced in the future.

4.1 Connection to the master program

The master program in Data Science and Advanced Analytics, as the name suggests, has a strong component on data analysis. This component, included a substantial practice of working with data sets using mainly SQL which was extremely important and useful. In fact, the work that the Model Unit Development does involves working with large data sets, in order to extract relevant information to develop predictive models. For instance, in the data cleaning processes it is necessary to analyse the existence and removal of outliers, dealing with missing values, or even removing duplicates, which were procedures already learnt and executed in the master program. These procedures had to be done in most of the projects where I was involved in order to have a cleaned data set to work with. Therefore, the practice of analysing and cleaning data learnt during the master program provided useful insights and knowledge which were essential to apply in the internship.

Most importantly, the master program involved working together in teams to develop specific projects. Team working experience was crucial to acquire the necessary skills for the professional world, such as, the ability of hearing others' opinions and reaching an agreement, as well as the capacity to organize and manage the tasks by all team members in order to meet deadlines with the required deliverables.

4.2 Internship Evaluation

One of the main goals of the internship, was particularly the development of forecasting models for ICAAP exercise. This project required a substantial effort and responsibility since it was a new methodology to put in practice and there were no existing procedures to follow in order to achieve its purpose.

However, I believe I worked extremely hard, giving my best to reach the best possible results. One of the key aspects which I believe was indispensable was starting to have an holistic view of all the tasks inherent to each phase of the projects having as much detail as possible and requesting the support of my supervisor whenever necessary. As a result, I started to put that idea in practice all the time noticing that it helps me a lot to be more conscious of what I am doing, for what purposes and what has to be done.

Overall, I have to come to realize that this type of work involves going backward and forward several times and that I need to accept that as being part of the process and absolutely normal. I also consider that, it helped me prepare to deal with my lack of confidence, and gave me the ability to deal optimistically with stressful and pressure working situations. Nonetheless, I believe that time and experience will help dealing with these situations and being more autonomous and confident.

4.3 Limitations

During the internship, in particular along the development of Balance Sheet projection models, there were several limitations, mainly regarding the data. As a matter of fact, it is very

different to work with real data since it requires working with large data sets which are not as clean and sometimes as useful, as the data used in university. Therefore, real life data is not as well behaved as it is in university and it becomes difficult to fulfil all statistical requirements which theoretically should always be met.

Additionally, the models implemented in the described project are considerably complex, requiring knowledge and technical understanding. Therefore, there is an additional challenge in explaining this statistical model to those who are not familiar with the subject. In particular, the customer/model owner may not have technical references to understand the model complexity.

Regarding the dimension of the team assigned to the project, at the beginning another colleague was also allocated to this project development, but then for personal reasons she left, so the responsibility of its development and implementation relied mainly on myself. Also, it relied on me to pass on all the acquired knowledge to the rest of the team, in particular to those which needed to apply this methodology for other entities of the group or even for other similar projects. Even though my supervisor was always available to help me, given the dimension and complexity of the project, combined with my lack of professional seniority, I believe that an additional member of the team should have been assigned for the project. Nonetheless, without this challenge I would not have acquired the skills and grown professionally so quickly as I did, so I am grateful it happened this way.

4.4 Lessons Learned

Along the internship, I was integrated in several projects with colleagues of the Development Model Unit team. The monitoring view of other models, which portrays different subjects, helped me learn about various key aspects of the banking business. For instance, learning how the scoring models work, how to estimate the losses of operational risk events, how the impairment of Loss Given Default may impact CGD's business, among others.

During the internship, not only technical skills were acquired, but also extremely important soft skills were developed. Namely, the ability of solving problems without being taught or having the necessary experience for it. Also, there was a remarkable improvement in working under pressure with established deadlines for projects deliveries, focusing in the actual problem without compromising the development and work deliverables. Working in teams also provides aptitudes in learning to hear others' opinions, and accepting other ways of thinking and proceeding. Especially, being able to identify personal traits in order to recognise characteristics of colleagues or superiors that can be useful to help with specific topics or issues was a crucial acquired competence. For instance, if having a doubt on a technical issue regarding the software used in a project, or wanting to know specific details about the banking business such as how the approval of a mortgage loan work, it is essential to identify who to turn to.

Thus, the internship was an extremely enriching experience where I developed a lot of my technical and personal skills. I would recommend anyone who is willing to start their professional career in a near future to do an internship in this scope. I believe that the skills

acquired in this kind of professional experience are priceless and unique, giving an adequate preparation for our professional careers.

4.5 Future Work

Regarding the Balance Sheet projection, a back testing exercise was performed in August by another colleague, corroborating the predictions obtained in the delivery of ICAAP exercise in March. Time series forecasting implemented in this project is the methodology aimed to use in the next ICAAP exercise, which reveals it is sufficiently robust to keep using for the time being.

Overall, I was in a friendly environment where I learned a lot with experienced professionals in the MDU, and in CGD in general, during the internship. Nevertheless, I believe there is still much more to learn in the banking sector, and for now, I was hired by CGD and will continue to make part of the team, doing my best to reach robust predictive models that assist in the management and decision-making of such a relevant banking institution as CGD.

To conclude, I believe that we should be ambitious and never settle for less, always seeking and working to learn more, that is, willing to do more and better. Effectively, we must try to continue to feel motivated and happy with our professional life. As the Chinese philosopher named Confucius once supposedly said: "Choose a job you love, and you will never have to work a day in your life".

Bibliography

[Agrawal et al., 2018] Agrawal, R. K., Muchahary, F., and Tripathi, M. M. (2018). Long term load forecasting with hourly predictions based on long-short-term-memory networks. In *2018 IEEE Texas Power and Energy Conference (TPEC)*. IEEE.

[Alkali et al., 2019] Alkali, M. A., Sipan, I. A. B., and Razali, M. N. (2019). The impact of macroeconomic variables on real estate price forecasting modelling in abuja nigeria. *International Journal of Engineering and Advanced Technology*.

[Almasarweh and Alwadi, 2018] Almasarweh, M. and Alwadi, S. (2018). Arima model in predicting banking stock market data. *Modern Applied Science*.

[Arunraj et al., 2016] Arunraj, N. S., Ahrens, D., and Fernandes, M. (2016). Application of sarimax model to forecast daily sales in food retail industry. *International Journal of Operations Research and Information Systems (IJORIS)*.

[Azari et al., 2019] Azari, A., Papapetrou, P., Denic, S., and Peters, G. (2019). Cellular traffic prediction and classification: A comparative evaluation of lstm and arima. In *International Conference on Discovery Science*. Springer.

[Barman, 2020] Barman, A. (2020). Time series analysis and forecasting of covid-19 cases using lstm and arima models. *arXiv preprint arXiv:2006.13852*.

[Bleikh and Young, 2016] Bleikh, H. Y. and Young, W. L. (2016). *Time series analysis and adjustment: Measuring, modelling and forecasting for business and economics*. CRC Press.

[Brocklebank and Dickey, 2003] Brocklebank, J. C. and Dickey, D. A. (2003). *SAS for forecasting time series*. John Wiley & Sons.

[Brockwell and Davis, 2009] Brockwell, P. J. and Davis, R. A. (2009). *Time series: theory and methods*. Springer Science & Business Media.

[Catalão et al., 2007] Catalão, J. P. d. S., Mariano, S. J. P. S., Mendes, V., and Ferreira, L. (2007). Short-term electricity prices forecasting in a competitive market: A neural network approach. *electric power systems research*.

[Chakraborty et al., 2019] Chakraborty, T., Chattopadhyay, S., and Ghosh, I. (2019). Forecasting dengue epidemics using a hybrid methodology. *Physica A: Statistical Mechanics and its Applications*.

[Chatfield, 2000] Chatfield, C. (2000). *Time-series forecasting*. CRC press.

[Cochrane, 2005] Cochrane, J. H. (2005). Time series for macroeconomics and finance. *Manuscript, University of Chicago*.

[Cools et al., 2009] Cools, M., Moons, E., and Wets, G. (2009). Investigating the variability in daily traffic counts through use of arimax and sarimax models: assessing the effect of holidays on two site locations. *Transportation research record*.

[Cowpertwait and Metcalfe, 2009] Cowpertwait, P. S. and Metcalfe, A. V. (2009). *Introductory time series with R*. Springer Science & Business Media.

[Cracan, 2020] Cracan, C. (2020). Retail sales forecasting using lstm and arima-lstm: A comparison with traditional econometric models and artificial neural networks.

[Davies et al., 2014] Davies, R., Coole, T., and Osipyw, D. (2014). The application of time series modelling and monte carlo simulation: Forecasting volatile inventory requirements. *Applied Mathematics*.

[De Gooijer and Hyndman, 2006] De Gooijer, J. G. and Hyndman, R. J. (2006). 25 years of time series forecasting. *International journal of forecasting*.

[Dinardi, 2020] Dinardi, F. B. (2020). *Forecasting the stock market using ARIMA and ARCH/GARCH approaches*. PhD thesis.

[Enders, 2008] Enders, W. (2008). *Applied econometric time series*. John Wiley & Sons.

[Engle, 1982] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the econometric society*.

[Frimpong and Oteng-Abayie, 2006] Frimpong, J. M. and Oteng-Abayie, E. F. (2006). Modelling and forecasting volatility of returns on the ghana stock exchange using garch models.

[GERUNOV, 2016] GERUNOV, A. A. (2016). Automating analytics: Forecasting time series in economics and business. *Journal of Economics and Political Economy*.

[Green, 2001] Green, A. J. (2001). Mass/length residuals: measures of body condition or generators of spurious results? *Ecology*.

[Ho et al., 2002] Ho, S.-L., Xie, M., and Goh, T. N. (2002). A comparative study of neural network and box-jenkins arima modeling in time series prediction. *Computers & Industrial Engineering*.

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*.

[Hyndman and Athanasopoulos, 2018] Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[Institute, 2012] Institute, S. (2012). *SAS/STAT 12.1 User's Guide*. SAS Institute Inc.

[Institute, 2014] Institute, S. (2014). *SAS/STAT 13.2 User's Guide The AUTOREG Procedure*. SAS Institute Inc.

[Kohzadi et al., 1996] Kohzadi, N., Boyd, M. S., Kermanshahi, B., and Kaastra, I. (1996). A comparison of artificial neural network and time series models for forecasting commodity prices. *Neurocomputing*.

[Lamoureux and Lastrapes, 1990] Lamoureux, C. G. and Lastrapes, W. D. (1990). Heteroskedasticity in stock return data: Volume versus garch effects. *The journal of finance*.

[Leszko, 2020] Leszko, D. (2020). *Time series forecasting for a call center in a Warsaw holding company*. PhD thesis.

[Liu et al., 2021] Liu, X., Lin, Z., and Feng, Z. (2021). Short-term offshore wind speed forecast by seasonal arima-a comparison against gru and lstm. *Energy*.

[McAleer and Da Veiga, 2008] McAleer, M. and Da Veiga, B. (2008). Forecasting value-at-risk with a parsimonious portfolio spillover garch (ps-garch) model. *Journal of Forecasting*.

[Mombeini and Yazdani-Chamzini, 2015] Mombeini, H. and Yazdani-Chamzini, A. (2015). Modeling gold price via artificial neural network. *Journal of Economics, business and Management*.

[Muzaffar and Afshari, 2019] Muzaffar, S. and Afshari, A. (2019). Short-term load forecasts using lstm networks. *Energy Procedia*.

[Namazian et al., 2018] Namazian, A., Ghodsi, M., and Nawaser, K. (2018). Prediction of temperature variations using artificial neural networks and arima model. *International Journal of Industrial and Systems Engineering*.

[Papacharalampous et al., 2018] Papacharalampous, G., Tyralis, H., and Koutsoyiannis, D. (2018). Predictability of monthly temperature and precipitation using automatic time series forecasting methods. *Acta Geophysica*.

[Prybutok et al., 2000] Prybutok, V. R., Yi, J., and Mitchell, D. (2000). Comparison of neural network models with arima and regression models for prediction of houston's daily maximum ozone concentrations. *European Journal of Operational Research*.

[Razali et al., 2011] Razali, N. M., Wah, Y. B., et al. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*.

[Safi, 2004] Safi, S. K. (2004). *The efficiency of OLS in the presence of auto-correlated disturbances in regression models*. American University.

[Sak et al., 2014] Sak, H., Senior, A. W., and Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

[Sjölander, 2011] Sjölander, P. (2011). A stationary unbiased finite sample arch-lm test procedure. *Applied Economics*.

[Sparks and Yurova, 2006] Sparks, J. and Yurova, Y. (2006). Comparative performance of arima and arch/garch models on time series of daily equity prices for large companies. *Unive Illinois Chicago*.

[Sulasikin et al., 2020] Sulasikin, A., Nugraha, Y., Kanggrawan, J., and Suherman, A. L. (2020). Forecasting for a data-driven policy using time series methods in handling covid-19 pandemic in jakarta. In *The 6th IEEE International Smart Cities Conference (ISC2 2020)*.

[Taylor, 2008] Taylor, S. J. (2008). *Modelling financial time series.* world scientific.

[Victor-Edema and Essi, 2016] Victor-Edema, U. A. and Essi, I. D. (2016). Autoregressive integrated moving average with exogenous variable (arimax) model for nigerian non-oil export. *European Journal of Business and Management.*

[Wang et al., 2005] Wang, W., Van Gelder, P., Vrijling, J., and Ma, J. (2005). Testing and modelling autoregressive conditional heteroskedasticity of streamflow processes. *Nonlinear processes in Geophysics.*

[Wu and Philip, 2005] Wu, E. H. and Philip, L. (2005). Volatility modelling of multivariate financial time series by using ica-garch models. In *International Conference on Intelligent Data Engineering and Automated Learning.* Springer.